

FUNDAÇÃO GETULIO VARGAS
ESCOLA DE ADMINISTRAÇÃO DE EMPRESAS DE SÃO PAULO

JOÃO GABRIEL SARACENI LIMA DA SILVA

NOVAS FONTES DE DADOS PARA INTELIGÊNCIA ANALÍTICA

SÃO PAULO
2018

JOÃO GABRIEL SARACENI LIMA DA SILVA

NOVAS FONTES DE DADOS PARA INTELIGÊNCIA ANALÍTICA

Dissertação apresentada à Escola de Administração de Empresas de São Paulo, da Fundação Getulio Vargas, como requisito para obtenção do título de Mestre em Administração de Empresas

Campo de conhecimento: Administração, Análise e Tecnologia de Informação

Orientador: Prof. Dr. Fernando de Souza Meirelles

SÃO PAULO
2018

Silva, João Gabriel Saraceni Lima da.

Novas fontes de dados para inteligência analítica / João Gabriel Saraceni Lima da Silva. - 2018.

95 f.

Orientador: Fernando de Souza Meirelles

Dissertação (mestrado) - Escola de Administração de Empresas de São Paulo.

1. Gestão do conhecimento. 2. Processo decisório - Processamento de dados. 3. Sistemas de suporte de decisão. I. Meirelles, Fernando de Souza. II. Dissertação (mestrado) - Escola de Administração de Empresas de São Paulo. III. Título.

CDU 65.012.4

JOÃO GABRIEL SARACENI LIMA DA SILVA

NOVAS FONTES DE DADOS PARA INTELIGÊNCIA ANALÍTICA

Dissertação apresentada à Escola de Administração de Empresas de São Paulo da Fundação Getúlio Vargas, como requisito para obtenção do título de Mestre em Administração de Empresas

Campo de conhecimento: Administração, Análise e Tecnologia de Informação.

Data de aprovação:

Banca Examinadora:

Prof. Dr. Fernando de Souza Meirelles
FGV-EAESP

Prof. Dr. José Luiz Carlos Kugler
FGV-EAESP

Prof. Dr. Gustavo Hermínio Salati
Marcondes de Moraes
UNICAMP

Dedico este trabalho a minha mãe.

AGRADECIMENTO

Agradeço a todo corpo docente da linha de pesquisa AATI, que me recebeu no programa de mestrado, e especialmente a meu orientador Professor Dr. Fernando de Souza Meirelles, por todo auxílio e condução durante a formação.

Ao amigo Alex da Silva Alves, que além da motivação para ingresso e persistência nessa jornada, contribui com seu companheirismo, boas conversas e verdade nas palavras.

A minha esposa pelo apoio e compreensão durante essa trajetória, e ao meu filho por ressignificar tudo e me trazer tão bons sentimentos.

Aos meus superiores imediatos nas empresas em que trabalhei, como Emerson Colin e Fabio Ono, que me ofereceram inspiração de carreira, boas referências profissionais e pessoais.

A todos amigos e familiares que contribuíram com apoio e boas energias, me ajudando a atravessar esse desafio.

RESUMO

A diversificação das fontes de dados utilizadas em processos decisórios nas organizações é um dos elementos que fundamentam o conceito de big data, apontado como o futuro das aplicações de Inteligência Analítica. O desafio das organizações em trabalhar com dados não estruturados e dados externos torna-se importante para as que desejam evoluir suas iniciativas de Inteligência Analítica. A partir da revisão de literatura e entrevistas com profissionais que atuam neste campo, este trabalho explora quais as principais iniciativas para obtenção de novas fontes de dados em sistemas de informação de Inteligência Analítica. A abordagem metodológica utilizada foi o estudo de caso múltiplo. A pesquisa utilizou como perspectiva de análise a Teoria da Capacidade Absortiva, que oferece elementos para avaliar como a empresa obtém informações externas a ela e as utiliza no contexto organizacional, gerando valor a seu negócio. São discutidos os principais desafios para a diversificação das fontes de dados, que pode se dar em função da complexidade de diferentes tecnologias, fornecedores, integração de dados, entre outros fatores. Os resultados são apresentados explorando os constructos de Aquisição, Assimilação, Transformação e Utilização, presentes na teoria da Capacidade Absortiva, aplicado à diversificação das fontes de dados nas organizações. Foi possível notar que existem diferentes níveis de utilização de dados externos nas organizações, bem como diferentes arranjos organizacionais para consumir dados externos. As diferentes formas de consumir dados externos causam impacto na forma de alocação de investimentos, governança de dados, cultura organizacional, relacionada ao uso do dado e, por fim, na maturidade do uso de dados externos de forma sistêmica na organização.

Palavras-chave: Inteligência Analítica, *Business Intelligence*, *Business Analytics*, *Big Data*, Dados externos, Dados não estruturados.

ABSTRACT

The variety of data sources applied into decision-making processes in organizations is one of the factors that defines the concept of big data, indicated as the future of Business Intelligence & Analytics applications. Handle with not structured and/or external data diversifying data sources became important to organizations that evolve their BI&A initiatives. From the literature review and interviews with field professionals, this work explore the main initiatives of organizations to obtain new data sources in BI&A information systems. The methodology approach used is multiple case study. The research used the Absorptive Capacity theory as analytical perspective, which offers elements to evaluate how organization obtains external information and use it into organizational context. There are discussed the main challenges to diversifying data sources, such as the complexity of new technologies, suppliers, data integration, among others. The results are presented exploring the constructs Acquisition, Assimilation, Transformation and Explotation, built in Absorptive Capacity theory, applied to the diversification of data sources. It was possible to notice that there are different external and/or not structured data usage levels in organization, as well as different organizational arrangements for consuming external data. The ways of consuming external data have effects on the investment allocation, data governance, organizational culture related to the use of data, and ultimately on the maturity of external and/or not structured data usage in a systematic way in the organization.

Keywords: Business Intelligence, Analytical Intelligence, *Big Data*, External data, Unstructured data.

LISTA DE FIGURAS

Figura 1: Framework de Business Intelligence	25
Figura 2: Framework de fontes de dados nas aplicações de Inteligência Analítica	26
Figura 3: Evolução do volume de dados estruturados e não-estruturados	28
Figura 4: Espectro de compartilhamento de dados.....	32
Figura 5: Capacidade Absortiva segundo Zahra e George (2002)	50
Figura 6: Capacidade Absortiva, segundo Todorova e Dursin (2007).....	52
Figura 7: Síntese dos principais aspectos relacionados a diversificação das fontes de dados..	82
Figura 8: Estágios de desenvolvimento da organização na diversificação de fontes de dados	83

LISTA DE TABELAS

Tabela 1: Tipos de fontes de dados	33
Tabela 2: Principais dimensões de diversidade de fontes de dados em sistemas de IA identificadas na literatura.....	48
Tabela 3: Dimensões da Capacidade Absortiva e suas principais características, segundo Zahra e George (2002).....	51
Tabela 4: Principais características dos entrevistados	56
Tabela 5: Roteiro construído para condução das entrevistas.....	57
Tabela 6: Utilização de codificação das entrevistas, e seu relacionamento com elementos teóricos	60

Sumário

1	Introdução	12
2	Revisão de literatura	14
2.1	Inteligência Analítica.....	14
2.2	Processo decisório baseado em dados	19
2.3	Diversidade de fontes de dados	22
2.3.1	Estrutura e armazenamento de dados	27
2.3.2	Formas de acesso aos dados	30
2.3.3	Métodos de obtenção de dados	41
2.4	Resumo da literatura	47
3	Referencial teórico	49
3.1	Teoria da Capacidade Absortiva.....	49
4	Método	54
5	Apresentação dos resultados.....	61
5.1	Aquisição	61
5.2	Assimilação.....	68
5.3	Transformação	71
5.4	Utilização.....	74
6	Conclusões	79
	Referências	85
	Anexo I	94
	Anexo II.....	95

1 INTRODUÇÃO

O debate sobre o tema de Inteligência Analítica tem crescido significativamente no ambiente acadêmico, bem como nas atividades profissionais aplicadas ao mercado. É crescente o número de publicações acadêmicas sobre o tema, que oferece diversas ramificações e possibilidades de debate (Chen, Chiang, & Storey, 2012). De forma similar, nos últimos anos, os investimentos em Inteligência Analítica (IA) no mercado têm figurado como principal preocupação dos CIOs. Em 2017 é indicada por 38% dos executivos estadunidenses como uma das três principais áreas de investimento de suas organizações (Struckman & Spencer, 2017). Adicionalmente, existe uma tentativa da indústria (principalmente por parte dos fornecedores) de justificar que investimentos em IA podem se converter em benefício econômico para as organizações. Por exemplo, num survey conduzido pelo *MIT Sloan Management Review* e *IBM Institute for Business Value*, com mais de 3.000 executivos de negócio em mais de 100 países, revelou-se que empresas com alto desempenho usam IA cinco vezes mais que as de baixa performance (LaValle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011).

Uma das vertentes de debate sobre IA que tem ganhado muita força é o tratamento de grandes volumes de dados, popularmente chamado de *big data*. A motivação para sustentar essa discussão se baseia no crescente volume de dados digitais disponíveis para consumo de usuários e empresas, por meio da popularização do uso da internet (ICT, 2016), bem como de seus serviços disponíveis. Este debate está estruturado não apenas no crescente volume de dados, mas também pela variedade de dados e diversificação de suas fontes, que abastecem os processos de gestão e análise da informação.

A importância da Inteligência Analítica, no contexto organizacional, fundamenta-se na necessidade de uso de informações como suporte ao processo decisório. Contudo, este uso de dados contrasta com a limitação das fontes das organizações, construindo um cenário próprio para o debate sobre alternativas disponíveis para acesso a novos dados. A maior disponibilidade de dados digitais constrói o atual paradigma, que combina dimensões como Volume, Veracidade, Velocidade, Variedade e Valor de dados, chamados de “5Vs”, para a caracterização do fenômeno *Big Data* (Demchenko, Grosso, de Laat, & Membrey, 2013). Para além do amadurecimento tecnológico experimentado pela sociedade, a mudança na forma como as organizações utilizam os dados, no auxílio de processos decisórios, muda importantes aspectos

de gestão, nos quais os dados assumem protagonismo entre os ativos de uma organização (Abbasi, Sarker, & Chiang, 2016; Redman, 2008).

De forma complementar a centralidade que as informações estão assumindo no ambiente de negócio, é fundamental que as organizações estejam atentas a necessidade de evolução dos sistemas de informação para dialogar com diferentes estruturas de dados, para além do ambiente interno das organizações (Ferrández, et al., 2016). As diferentes fontes de dados diferem em estrutura, modelagem e padrões tecnológicos, num “ecossistema” cada vez mais rico e complexo. Atualmente, é comum que nem todas as informações da organização estejam em repositórios de seu completo domínio, ficando fisicamente armazenadas em bancos de dados de serviços contratados na web (como ações de marketing digital, por exemplo), hospedados em servidores em nuvem, ou totalmente dispersas em diferentes domínios na internet.

Os desafios de interação da empresa com essa diversidade de fontes de dados constituem o escopo do presente trabalho. Nesse sentido, o objetivo da pesquisa é explorar quais características associadas à iniciativas para obtenção de novas fontes de em sistemas de informação de Inteligência Analítica. Deste modo, a contribuição prática desta pesquisa se dá pela identificação de elementos que gestores de projetos de Inteligência Analítica terão de superar, caso tenham intenção de iniciar um processo de diversificação de fontes de dados em sua organização. No âmbito teórico essa pesquisa contribui na medida que explora questões relacionadas a diversificação de fontes de dados, elemento relevante na literatura relacionada a Inteligência Analítica, embora pouco explorado.

2 REVISÃO DE LITERATURA

O campo de estudo de Inteligência Analítica possui um relevante número de publicações acadêmicas e alvo de muitas discussões profissionais relacionadas à suas práticas nas organizações. Este trabalho se dedicará a apresentar uma revisão de literatura direcionada à discussão sobre os principais elementos diretamente relacionados ao fenômeno estudado. Assim, a revisão se inicia pelo delineamento de escopo do que se refere inteligência analítica, bem como a apresentação de seus principais conceitos e referências. Em seguida é apresentado uma revisão sobre a visão de processos decisórios baseados em dados, e sua crescente importância nas organizações. Na sequência é apresentada uma visão sobre a diversificação de dados em projetos de inteligência analítica. Essa revisão foi organizada de modo a estabelecer os principais elementos relacionados a estrutura e armazenamento de dados; as formas de acesso aos dados e; os métodos de obtenção de dados possíveis de serem utilizados atualmente. Este item se encerra apresentando um breve resumo da literatura, que deverá suportar a construção do instrumento de coleta utilizado na presente pesquisa.

2.1 **Inteligência Analítica**

A definição de Inteligência Analítica evoluiu ao longo de décadas, variando sua abrangência em termos de processos de negócios, manipulação de dados, envolvendo competências de pessoas e de organizações. Do ponto de vista acadêmico e profissional, este é um campo de estudo que pode sofrer a influência de muitos rótulos, tais como *Business Intelligence* (BI), *Business Analytics* (BA), *Big Data*, *Decision Support Systems* (DSS), entre outros. Meirelles (2017), em sua pesquisa anual de Administração e Uso da TI, com mais de 2.500 empresas participantes no Brasil, destaca que o termo *Business Intelligence* é tão utilizado comercialmente que há dificuldade em entender realmente que fenômeno ele descreve. O autor considera o termo Inteligência Analítica (IA) como substituto compatível para o contexto acadêmico, principalmente pelo fato de não ser utilizado comercialmente. Alinhado com este conceito, este termo é adotado para o desenvolvimento do presente trabalho.

Sob o ponto de vista acadêmico, este campo de pesquisa é relativamente antigo. Por exemplo, em 1958 Luhn publicou sua visão sobre o que seria um sistema de informação batizado de *Business Intelligence System*. No trabalho, BI foi conceituado como um sistema automático capaz de sugerir possíveis ações reais para os usuários, partindo sempre de documentos e dados. Luhn (1958) descreve de forma técnica o que esse sistema deve fazer, mas não são citados possíveis benefícios e nem para onde se esperava que o BI fosse se desenvolver. Anos depois, Gilad e Gilad (1986) definiram *Business Intelligence* como um processo no qual a entrada são os dados brutos e a saída é inteligência para tomada de decisão.

Outros autores confirmam o conceito de *Business Intelligence* como um processo, que utiliza informações operacionais e ferramentas analíticas para melhorar os processos de decisão das empresas (Negash, 2004). Desde o surgimento do conceito de BI, o campo de estudos experimentou constantes evoluções, sobretudo em função da evolução do contexto em que ele se insere, tornando-o mais poderoso. Por exemplo, a capacidade de tratamento de dados, bem como dos softwares, hardwares e sistemas de informação, evoluiu significativamente, possibilitando que as ferramentas de BI aumentassem seu alcance a diferentes funções da organização.

É oportuno observar que o texto de Gilad e Gilad (1986) aponta BI como um processo, mas orienta-se a uma abordagem de usos de dados oriundos do ambiente externo à empresa. Durante alguns anos essa atividade era particularmente complexa, devido à dificuldade de acesso aos dados de forma legível, por máquinas, de modo que os sistemas de IA se desenvolveram basicamente a partir de dados internos da organização, mais alinhados à definição de Negash (2004). Hoje, com a modernização dos sistemas de informação e sua popularização na sociedade, a possibilidade de trabalhar dados externos aumenta sensivelmente, ajudando as ferramentas de IA a se reaproximarem das definições de Gilad e Gilad (1986).

Alguns autores definem BI não apenas como um processo, mas também como um produto (Vedder, Vanecek, Guynes, & Cappel, 1999). Para eles, o processo é composto de métodos para tratar a informação que as empresas podem usar para competir. O produto é a informação que permite predição de comportamento de competidores, fornecedores, clientes, tecnologias, entre outros, com determinado grau de acerto. De forma complementar, Bose (2009) alega que BI pode ser definido sob duas perspectivas. A primeira é gerencial, na qual BI pode ser definido como uma forma de entregar as informações corretas, para as pessoas

corretas, para que tomem decisões e melhorem o desempenho do negócio. A segunda perspectiva tem caráter mais técnico, definindo BI como um conjunto de tecnologias para reunir, armazenar, analisar e prover acesso a dados como uma forma de melhorar a tomada de decisão.

Numa perspectiva histórica, é possível identificar que o processo de evolução das ferramentas de IA se deu de forma gradual. Na década de 70 surgiram pacotes de *software* analíticos no mercado, como as planilhas, por exemplo. Já década de 80 surgiram os primeiros sistemas de informação com foco executivo e os DDS (*Decision Support Systems*). Eles permitiam rápido acesso a informações internas, mas ao longo dos anos provaram dar uma visão muito estreita para a tomada de decisão. Na década de 90, surgiram os sistemas de BI, uma soma de *Data Warehouse*, ferramentas de ETL (*Extract, Transform, Load*), e *softwares* analíticos, com capacidade OLAP (*Online Analytical Processing*) (Petrini & Pozzebon, 2009).

Originalmente, Inteligência Analítica baseava-se em grandes quantidades de dados estruturados, geralmente na forma de *Data Warehouse* (Holsapple, Lee-Post, & Pakath, 2014), que pode ser definido como uma coleção de dados utilizada para fins de tomada de decisão (Baars & Kemper, 2008). Um *Data Warehouse* é constituído de dados de diversas fontes. Assim, operações de extração, transformação e carga (ETL) são necessárias. Para tratar os dados, surgiram novas tecnologias como o *Hadoop* e o *MapReduce*, que têm o objetivo final de melhorar ganhos de conhecimento, gerar *insights*, encontrar e resolver problemas, como forma de apoiar decisões (Holsapple, Lee-Post, & Pakath, 2014). As informações consideradas podem ser quantitativas e também qualitativas, uma vez que ambos os tipos interessam aos tomadores de decisão.

O contexto de aplicação de ferramentas de BI tem evoluído, sendo um dos aspectos o aumento da disponibilidade de fontes de dados para contribuir ao processo decisório. Para além do processo de informatização das organizações, outra variável importante é a evolução da infraestrutura de internet e o aumento de usuários no mundo. Para além do amadurecimento tecnológico experimentado pela sociedade, a mudança na forma como as organizações utilizam os dados, inserindo-os no auxílio de processos decisórios, altera importantes aspectos de gestão, assumindo um protagonismo entre os ativos da organização (Redman, 2008).

O crescimento de disponibilidade de dados digitais oferece às empresas um conjunto de novas possibilidades para análise de seus negócios. A quantidade de trabalhos acadêmicos que abordam o tema acompanha esse crescimento, discutindo possibilidades de aplicações dos

sistemas de informação e do uso de dados no apoio ao processo decisório nas organizações. Esse aumento de importância dada ao tema criou um ciclo de grande popularidade ao termo *big data*, que sintetizou diversas iniciativas relacionadas à gestão da informação, tornando-se um termo de difícil decisão e compreensão.

As definições acerca do termo *big data* podem variar no que tange a que dimensões deste fenômeno devem ser consideradas para sua definição. Um conjunto dessas dimensões, também aceito e utilizado, considera aspectos como volume, velocidade, variedade, valor e veracidade (Demchenko, Grosso, de Laat, & Membrey, 2013). Essas dimensões podem ser utilizadas para descrever os requisitos e o que se deseja de tal fenômeno, o que, de certa forma, pode estar vinculado a expectativas comerciais relacionadas a cada dimensão. Na dimensão de volume de dados (dimensão de maior ênfase na origem do debate acerca deste termo) são apontadas estatísticas que buscam justificar sua importância. Em 2011, o volume de dados passível de utilização por técnicas de IA na internet estaria próximo a um zettabyte (2^{70} bytes), medido por meio do tráfego de dados, possivelmente atingindo a escalada de yottabyte (que equivale a 1000 zettabytes), antes do final do século (Kiron & Shockley, 2011). Outros autores citam que o volume de dados produzido por dia supera 2,5 exabytes, e que 90% dos dados disponíveis hoje foram produzidos nos últimos dois anos (Olmedilla, Martínez-Torres, & Toral, 2016)

Ainda no contexto de *big data*, Chen, Chiang e Storey (2012) adicionam o termo *Analytics* e definem *Big Data Analytics* como um termo para descrever conjuntos de dados e técnicas analíticas para aplicações muito grandes e complexas, que requerem soluções avançadas de armazenagem, gerenciamento, análise e ferramentas de visualização. Davenport (2006) definiu *Business Analytics* (BA) como o termo adequado para *Big Data Analytics* no contexto de negócio. Para esse autor, BA é o componente analítico do BI, que surgiu pela crescente complexidade desse componente na presença do fenômeno *big data*. Essa mistura de termos e conceitos é uma das principais características da área, o que causa divergências entre pesquisadores, praticantes e pesquisadores.

Chen, Chiang e Storey (2012), frente a essa mistura de terminologias, adotaram BI&A (*Business Intelligence and Analytics*), que se refere ao conjunto de técnicas, tecnologias, sistemas, práticas, metodologias e aplicações para analisar dados críticos de negócio. Este tem o objetivo de ajudar empresas a melhor entender o negócio, o mercado e realizar decisões em tempo adequado. Apesar da tentativa de consolidar um termo que unifique os conceitos, a

maioria dos autores continua utilizando diferentes terminologias para abordar o tema. Por isso, esse trabalho adota o termo Inteligência Analítica, cuja utilização não é comercial. A diferença entre as aplicações clássicas de BI e o atual IA não é apenas composta de volume, mas também da profundidade das informações e de que tipos de perguntas podem-se responder (Debortoli, Müller, & Vom Brocke, 2014).

Ao estabelecer uma relação entre as aplicações de *Business Intelligence* e a diversidade de fontes de dados, Chen et al. (2012) sugere que existem 3 grandes estágios de utilização dessas aplicações. O primeiro, denominado BI&A 1.0, refere-se a aplicações limitadas à análise de dados disponíveis em repositórios de dados tradicionais, tais como bancos de dados relacionais e dados internos à organização. No segundo estágio, denominado BI&A 2.0, dados disponíveis na Web e dados não estruturados passam a ser considerados na arquitetura de informações. Já no terceiro estágio, denominado BI&A 3.0, estão presentes dados gerados por máquinas, que passariam também a integrar a arquitetura, considerando, por exemplo, o ambiente de Internet das Coisas (IoT). A partir da definição de Chen et al. (2012), este trabalho se dedicará, principalmente, ao debate do segundo estágio proposto pelo autor.

O estágio definido como BI&A 2.0 deve-se, sobretudo, à evolução da internet, principalmente na década de 2000. A ascensão dos motores de busca e dos ambientes de comércio eletrônico permitiu que as empresas apresentassem seus negócios e interagissem com seus consumidores de forma direta no ambiente digital. Enquanto as empresas disponibilizam informações de seus produtos e do negócio na internet, a coleta de informações que usuários disponibilizam na internet se tornou um ativo importante a ser explorado pelos negócios. As novas possibilidades de obtenção de dados e realização de análise constitui a base do conceito associado ao BI&A 2.0, proposto por (Chen, Chiang, & Storey, 2012).

Outro aspecto que enriquece o debate sobre esse estágio refere-se ao potencial, represado nas organizações, de tratamento de dados tradicionalmente classificados como “não estruturados”. Estes dados, normalmente contidos em relatos de clientes (reclamações, registros de atendimentos, sugestões), trocas de e-mails e registros similares são de propriedade da empresa, sem que muitas vezes a mesma seja capaz de fazer um tratamento sistemático desses dados. Contudo, existe atualmente uma grande expectativa de que as ferramentas de análises sejam capazes de dar vazão a esse potencial armazenado nas organizações, posicionando o estágio de BI&A 2.0 como o de maior potencial de evolução dos sistemas de Inteligência Analítica.

2.2 Processo decisório baseado em dados

O debate sobre a importância do uso de dados como auxílio ao processo decisório é amplo, com diferentes posicionamentos. No passado mais recente, o debate sobre o uso mais estratégico de sistemas de informação ganhou importância, dando início à literatura de sistemas de suporte à decisão e similares. Este amadurecimento tem também relação com a maior possibilidade de sistemas processarem dados empresariais. Contudo, informações de nível tático e estratégico por vezes estão para além das fronteiras da organização (Ferrández, et al., 2016). Assim, a variedade de fontes de informação que a organização precisa conjugar para melhor suportar seus processos decisórios tende a crescer cada vez mais.

Apesar de o contexto contemporâneo oferecer uma maior disponibilidade de dados digitalizados, os desafios das organizações estão associados à sua capacidade de conjugar essas múltiplas e numerosas fontes de dados, de modo a gerar informações úteis para o processo decisório. Hal Varian, economista chefe na Google e professor emérito na Universidade Berkley (Califórnia-EUA), defende que essa imensa disponibilidade de dados se constitui numa oportunidade para essa geração de novos profissionais. Se os dados se tornam ubíquos e baratos, a análise desses dados é o instrumento complementar nesse cenário para alavancar a carreira profissional de quem atua neste mercado (Chen, Chiang, & Storey, 2012).

A adequação das organizações frente aos avanços, no contexto da tecnologia da informação, apresenta reflexos na incorporação destas tecnologias, nos níveis mais elevados de gestão. Para Power (2008), o termo *Business Intelligence* se popularizou, sendo utilizado para descrever um conjunto de conceitos e métodos para melhorar a tomada de decisão dos negócios, usando sistemas de informação baseados em fatos. Assim, a década de 80 marca uma importante mudança na visão da gestão sobre os sistemas de informação, colocando o uso de dados e fatos como elemento de suporte ao processo decisório. Ainda, ele originalmente se suporta no conceito de OLAP (*Online Analytical Processing*), que tem como um de seus fundamentos a capacidade de diversas fontes de dados.

Para a evolução de uma iniciativa de IA, um conjunto de fatores precisa ser trabalhado, sob o ponto de vista de gestão. Gudfinnsson et al. (2015) argumenta que um dos principais obstáculos a ser superado pelas empresas pode ser encontrado na falta de supervisão dos executivos e da cultura organizacional, no adequado encorajamento do compartilhamento da

informação. Adicionalmente, ele argumenta que o investimento em soluções de IA é muito subestimado e a própria duração de tempo destes projetos acaba por levantar importantes problemas, que precisam ser gerenciados, se a organização quiser alcançar os benefícios contabilizados pela literatura e praticantes. Acima de tudo, o CIO precisa exaltar constantemente a “*cultura analítica*”, quando for questionado sobre quais pontos críticos para o sucesso para iniciativas dessa natureza. Na mesma linha de pensamento, Davenport (2010) argumenta que a alta gestão das organizações deve buscar transparência nas informações, pois essa postura afeta a cultura da empresa, de modo a encorajar uma orientação analítica nos colaboradores.

Para reforçar a importância do suporte de dados ao processo decisório, McAfee e Brynjolfsson (2012) partem da afirmação de que “não se pode gerenciar aquilo que não é possível medir”. De acordo com esses autores, o ditado que tem sido atribuído a ambos, W. Edwards Deming e Peter Drucker, explica porque a recente explosão de dados digitais é tão importante. Eles sustentam que, por conta do fenômeno *big data*, os gestores podem medir e conhecer muito mais seus negócios e traduzir diretamente esse conhecimento na melhora do processo decisório e do *desempenho*. Um dos principais aspectos deste fenômeno é seu impacto sobre como as decisões são tomadas e de como construí-las. Quando dados são escassos, com custos altos de obtenção, ou não estão disponíveis na forma digital, faz sentido destacar o papel das pessoas no processo decisório. Essas pessoas se utilizam de suas experiências criadas, baseadas em padrões e relações que foram observadas e internalizadas. Normalmente é atribuído o rótulo de “intuição” a este tipo de inferência e de processo decisório. Assim, pessoas fazem afirmações a partir de suas opiniões sobre o que acontecerá no futuro, sobre o quanto algo dará certo, e assim por diante, fazendo planos de acordo com essas perspectivas.

Quando as decisões são particularmente importantes, essas pessoas estão tipicamente em altos níveis da organização, ou são consultores externos caros, trazidos para o processo decisório em função de sua expertise e experiências anteriores. Muitas empresas na comunidade de *big data* ainda tomam boa parte das decisões importantes baseadas na opinião das pessoas que mais ganham na empresa (“*HiPPO*”—*the highest-paid person’s opinion*) (McAfee e Brynjolfsson, 2012). De forma complementar, Davenport (2010) argumenta que tomar decisões baseadas no poder e na política, e não em análises objetivas, é como um câncer na cultura organizacional. Sugere ainda que, se você for poderoso o bastante na hierarquia, conseguirá o que deseja, mas isso nem sempre pode ser o melhor para a empresa.

Seguramente um grande número de executivos seniores está genuinamente direcionado por dados e dispostos a substituir sua própria intuição quando os dados apontam outras realidades. Apesar disso, acredita-se ainda que os negócios se baseiam muito na experiência e intuição, e não o suficiente em dados. McAfee e Brynjolfsson (2012) afirmam que, em suas pesquisas, 32% dos respondentes estão abaixo de três pontos, numa escala de cinco, sobre o quanto suas organizações são orientadas por dados. Davenport (2010) advoga ser muito difícil mudar uma cultura não analítica, pois é preciso superar os gerentes seniores que baseiam decisões importantes na intuição. Defende ainda a ideia de que mesmo a criação de um novo departamento de profissionais analíticos não é suficiente para mudar uma cultura não analítica.

Para McAfee e Brynjolfsson (2012), existe uma mudança fundamental a que se deve estar atento. Esta diz respeito à orientação da empresa em se questionar sobre “o que ela sabe”, ao invés de “o que ela acha”. Isso requer que ela se afaste das ações baseadas em palpites e instintos. Também significa quebrar um mau hábito notado em muitas organizações: fingir que estão mais orientadas por dados do que realmente estão. Com muita frequência, executivos incrementam seus relatórios com muitos dados para justificar suas decisões, tomadas utilizando métodos mais tradicionais, como o da “opinião da pessoa mais bem paga” (HiPPOs).

Na visão de Gudfinnsson et al. (2015), apesar de muitos gerentes seniores usarem soluções de BI, isso não significa que a empresa tem conseguido criar uma cultura analítica, ou incorporado efetivamente a análise de dados à estratégia de negócio. Em geral, os usos parecem focar mais num olhar pelo retrovisor e observar o que aconteceu, ao invés de olhar adiante, utilizando análise de dados para guiar decisões de negócio.

Os desafios de criar uma cultura analítica e verdadeiramente orientada por dados são complexos e necessitam de particular atenção da organização. Davenport (2010) defende que uma cultura analítica é uma meritocracia que reconhece e premia os melhores *insights* orientados por dados. Sendo assim, os executivos nas organizações devem entender as raízes da cultura analítica e buscar incorporá-las a seu cotidiano. Segundo o autor, culturas “não analíticas” usam histórias e relatos sem comprovação para sustentar suas decisões; as culturas analíticas buscam dados.

A migração para a cultura analítica pode fazer emergir um questionamento fundamental para a alta gestão sobre o excesso de apego às análises de dados, com aumento da burocracia em determinados processos e coletas de dados em excesso, impactando na perda de eficiência do processo decisório. Davenport (2010) argumenta que os executivos devem buscar equilíbrio

e racionalidade para diferenciar sobre quando é razoável ter disponibilidade adequada de dados, e quando a disponibilidade de dados é dispensável. Como regra geral, quão mais sensível é o impacto da decisão, mais adequado seria a preocupação para a obtenção de dados para suportar o processo decisório.

Davenport e Harris (2007) discorrem sobre características de recursos humanos que devem ser buscadas e qual o perfil que deve ser almejado ou estimulado em executivos nas empresas. Assim, os autores apresentam pontos importantes a serem trabalhados pelos executivos nas organizações, a fim de estimular uma cultura analítica e verdadeiramente orientada por dados. Eles destacam as principais características de um CEO:

- Deve ser incentivador apaixonado do processo de tomada de decisão analítico e baseado em fatos;
- Deve ter algum apreço por ferramentas e métodos analíticos;
- Deve desejar realizar ações a partir dos resultados das análises;
- Deve estar disposto a gerenciar em função da meritocracia.

2.3 Diversidade de fontes de dados

O fenômeno de aumento da disponibilidade de dados digitais, usualmente chamado de *big data*, tem como perspectiva de análise algumas dimensões tais como Volume, Veracidade, Velocidade e Variedade de dados, chamados de “4Vs” (Goes, 2014). A definição conceitual e a quantidade de dimensões de análise variam, dependendo do autor ou contexto. Demchenko, Grosso, de Laat, e Membrey (2013), por exemplo, consideram também a dimensão de Valor dos dados, formando então 5Vs. A visão corrobora com Kugler (2013), que advoga sobre a necessidade de a organização estabelecer “questões críticas do negócio”, com o objetivo de modelar a sua estrutura analítica e potencializar a geração de valor a partir dos dados. O presente trabalho considera essas dimensões como plano de fundo para discussão do impacto dos dados nos sistemas de Inteligência Analítica, mas explora especialmente a dimensão Variedade. Esta dimensão representa um desafio particular, que diz respeito à identificação, obtenção e conjugação de diferentes fontes de dados.

Esse conjunto diverso de fontes de dados tem sua tipificação variando em função de sua estrutura de dados e a forma de acesso/disponibilidade. As diferentes combinações dessas variáveis aumentam significativamente a complexidade na gestão de dados, a fim de aumentar sua utilização no suporte aos processos decisórios. Apesar da complexidade, o princípio de conjugação de múltiplas fontes de dados é relativamente antigo, dado que a essência do conceito de *Data Warehouse* (DW) é a integração de dados a partir de diferentes fontes, num único repositório de informações (Power, 2008). Atingir um nível satisfatório de integração entre dados internos e externos é um dos grandes desafios enfrentados pelo designer de DW. Por sua vez, a integração de dados está também no coração das ferramentas de suporte a decisão sustentada por DW. A competência de usar ferramentas de IA para suportar decisões táticas ou estratégicas depende de adequada integração de dados brutos, atuais e históricos, que possam impactar nas decisões (March & Hevner, 2005).

Na busca pela diversificação de fontes de dados em projetos de DW, alternativas tecnológicas são avaliadas como possibilidades para expansão do universo de dados de forma sistemática. Diferentes conceitos são considerados para a dinamização de fontes dados, tais como *Querying Knowledge Bases*, *Querying Data Management Platforms* e *Querying Semantic Web Search Engines* (Ibragimov, Hose, Pedersen, & Zimanyi, 2015).

As competências exigidas por profissionais responsáveis pela análise de dados compreendem uma diversidade de conhecimentos, entre eles a capacidade de manipulação de dados. Além da capacidade do indivíduo em obter dados de bases estruturadas, é necessário também a habilidade de obter dados de fontes alternativas, como, por exemplo, de mídias sociais como o Facebook, e transformá-los de modo a integrar dados internos e externos num único repositório (Schoenherr & Speier-Peró, 2015).

A importância do uso de dados externos à organização é reforçada por estudos como os de Kimble e Milolidakis (2015) que destacam o aumento da disponibilidade de dados na internet de forma significativa. O acesso a dados na internet é facilitado por serviços de streaming de dados, bem como a popularização de APIs (*Application Programming Interfaces*) e/ou aumento de tecnologias capazes de absorver dados de *websites*. A capacidade de acessar esses dados representa a possibilidade de obter maior volume de informações sobre clientes (atuais e potenciais), mercado e competidores. Assim, empresas estão cada vez mais interessadas em coletar dados externos (como mídias sociais, economia, entre outros) para enriquecer as fontes internas de dados, onde se baseiam tradicionalmente. Ferrández, et al. (2016) afirmam haver

um grande consenso de que dados internos às organizações não são suficientes para um processo decisório adequado, principalmente em mercados de grandes mudanças e dinamicidade, nos quais informações sobre clientes e competidores são extremamente relevantes para essas decisões. De forma complementar, Rodrigues (2017) chama atenção para o valor do uso de dados “alternativos” (obtidos no ambiente externo à empresa, e que não são solicitados a um cliente convencionalmente) nos processos anti-fraude, como um exemplo da utilização de dados externos.

Apesar de uma necessidade contemporânea e comum, a gestão das organizações é marcada pela dificuldade de integração de dados externos e internos. De acordo com Randall e Beyer (2015), pesquisadores do Gartner Group, o contexto atual das empresas é de dar mais atenção à exploração de dados internos, apesar de tentativas de diversificação de suas fontes de dados. Estes autores indicam a possibilidade de convergência das competências de tratamento dos dados externos no futuro. Mas, atualmente, as organizações têm se dedicado mais ao tratamento de dados internos.

Outro elemento que influencia no tratamento de fontes externas de dados é a função de Inteligência Competitiva (IC). Apesar da semelhança dos termos (em inglês *Competitive Intelligence* e *Business Intelligence*), Baars e Kemper (2008) defendem que cada área se desenvolveu de forma independente. Até recentemente, profissionais e pesquisadores de IC não exploraram muito o suporte de infraestrutura tecnológica, principalmente em função da heterogeneidade e natureza da maioria das fontes de dados (pesquisas em formatos de livros, convenções e conferências, *Surveys* em *Delphi*, ou bases de dados de patentes, por exemplo). Com o crescimento da importância da internet e de documentos eletrônicos, uma tendência da integração da função de IC com infraestrutura tecnológica pode ser observada. Assim, IC atualmente pode ser visto como uma aplicação no domínio de Inteligência Analítica. Deste modo, os autores apresentam uma visão ampliada de fluxo de informação, partindo dos processos organizacionais e de dados externos, onde os dados passam a ser trabalhados por uma camada de dados, uma camada lógica e uma camada de acesso ao produto de inteligência, conforme descrito na Figura 1 a seguir.

Para além da origem dos dados (internos ou externos), outro aspecto de fundamental importância está relacionado à estrutura de dados. É preciso considerar que uma fração relevante dos dados disponíveis está disponível de forma “não convencional”, por vezes chamados de dados “não estruturados” ou nomenclaturas similares. Apesar de o tema estar

ganhando importância na gestão de dados contemporânea, as organizações são relativamente incipientes nesta questão. De acordo com uma pesquisa da IDG Enterprise (2016), 17% das organizações afirmam considerar o tratamento de dados não estruturados como uma questão prioritária na organização, enquanto 45% consideram esta questão um grande desafio e apenas 31% acreditam ter esse problema sob controle. Os demais não consideram o tema pertinente para a empresa.

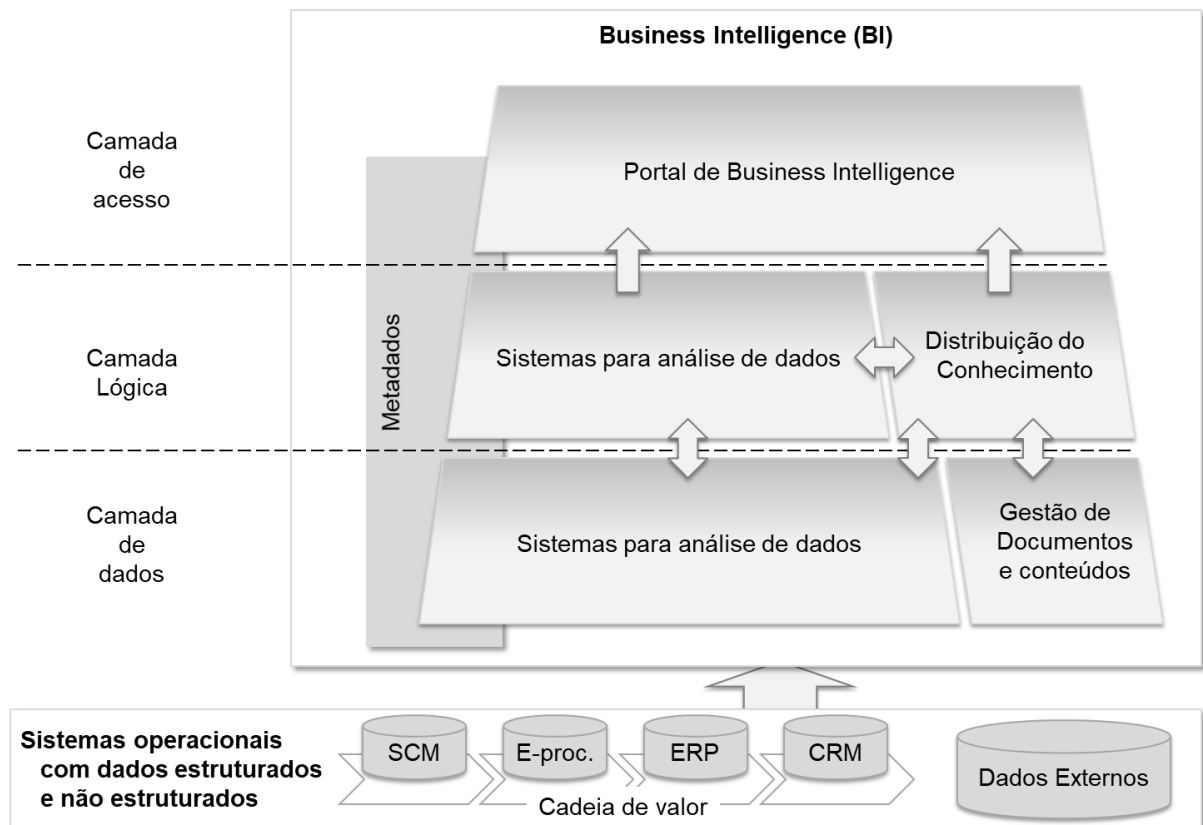


Figura 1: Framework de Business Intelligence
 Fonte: Baars e Kemper (2008), traduzido pelo autor

Baars e Kemper (2008) argumentam que o tratamento de dados não estruturados é uma questão chave para os sistemas de IA, dado que a maior parte das informações é oriunda de dados não estruturados ou semiestruturados, como exemplo, e-mails de clientes, páginas da web com informações dos competidores, relatórios de vendas, repositórios de artigos de pesquisa, entre outros. Especialmente em áreas de alcance além das fronteiras da organização, como Gestão de Relacionamento com Consumidor (*Customer Relationship Management - CRM*) ou Inteligência Competitiva, o tratamento de dados estruturados e não estruturados torna-se imperativo.

Como mencionado anteriormente, o trabalho de Chen et al. (2012) sugere a classificação de três estágios de utilização de BI&A. O primeiro estágio, que contempla o uso mais tradicional da técnica, é denominado como BI&A 1.0. A ampliação da utilização das técnicas aos dados disponíveis na Web, como BI&A 2.0. Por fim, a utilização de informações oriundas de dispositivos IoT e Mobile como BI&A 3.0.

A partir do trabalho de Chen (2012) é possível estabelecer uma relação entre as fontes com as variáveis de estrutura e acessibilidade dos dados. Sendo assim, é possível perceber que futuros projetos de Inteligência Analítica tenderão a utilizar cada vez mais dados não estruturados e/ou dados para além dos limites de controle da empresa (externos à organização). O diagrama da Figura 2 a seguir evidencia a abrangência das fontes de dados que deverão ser incorporadas a projetos de Inteligência Analítica.

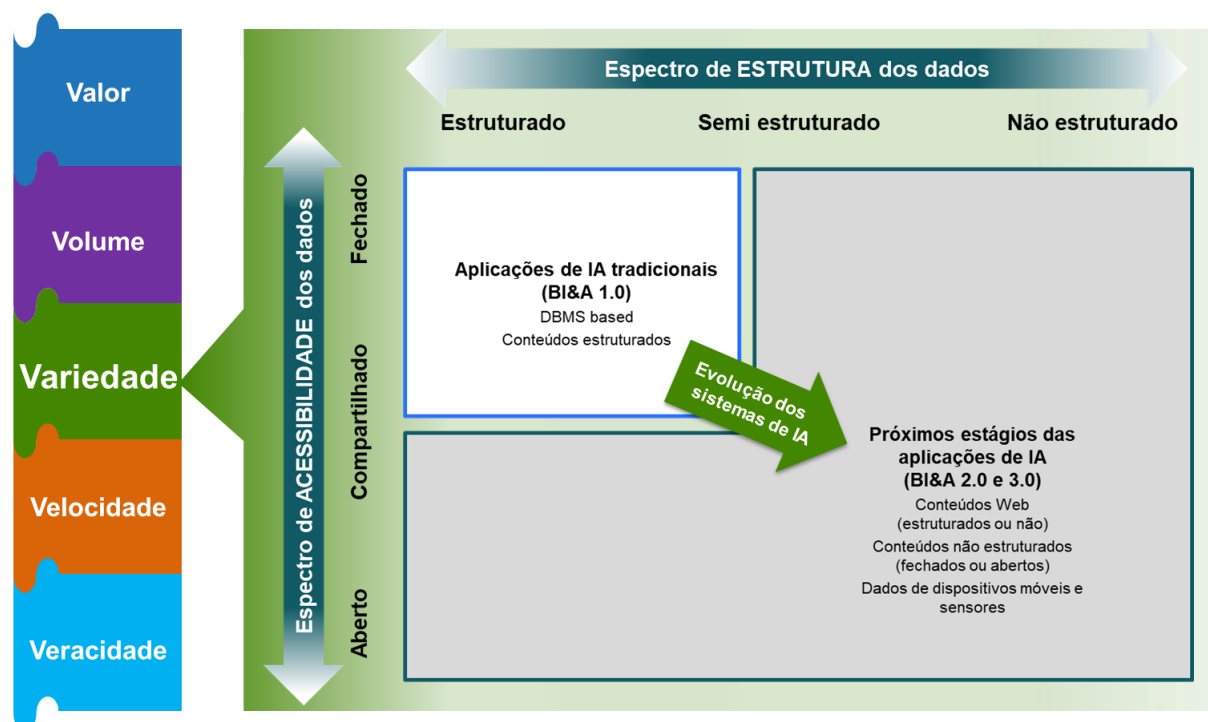


Figura 2: Framework de fontes de dados nas aplicações de Inteligência Analítica
 Fonte: Baseado em Negash (2004) – Data Types/Source Matrix; Chen, et al. (2012) e Demchenko, Y., Grosso, P., Laat, C., e Membrey, P. (2013)

A utilização de diferentes estruturas/modelos de dados, bem como fontes de origem externa, são dimensões distintas, embora possam ser trabalhadas de forma conjugada. Os próximos itens deste trabalho exploram essas dimensões de forma detalhada, descrevendo características relacionadas a iniciativas de Inteligência Analítica no contexto organizacional. No item 2.3.1 são apresentados elementos relacionados à importância do olhar das empresas

sobre a utilização de dados em estruturas não convencionais, ditos como dados não estruturados. No item 2.3.2 a dimensão de acessibilidade é aprofundada, apresentando diferentes formatos de compartilhamento de dados e agentes que atuam nesta dimensão.

2.3.1 Estrutura e armazenamento de dados

Convencionalmente a literatura trata a discussão da estrutura de dados dividindo-os em dados estruturados, semiestruturados e não estruturados. Contudo, definir cada um desses grupos torna-se uma tarefa complexa, em função do conceito difuso sobre o que é um dado estruturado. Esta reflexão acompanha o desenvolvimento de sistemas de gestão de banco de dados em toda a história e as convenções sobre o tema surgem em paralelo com a consolidação dos princípios de bancos de dados relacionais e a linguagem de consulta estruturada (SQL - *Structured Query Language*) (Berg, Seymour, & Goel, 2013 ; Brackett, 2011).

Sendo assim, o conceito de dados “não estruturados” pode ser amplamente debatido, pois os dados, usualmente, possuem uma estrutura, por mais inconveniente que seja o seu tratamento por máquinas. Sullivan (2004) afirma que esse termo é geralmente mal utilizado, uma vez que linguistas possam afirmar que textos possuem uma estrutura bastante clara e definida. Nesse gradiente de estrutura de dados, existem propostas do uso de terminologias como dados multi-estruturados, ou poliestruturados, para se referir a dados com estruturas mais complexas (semi-estruturados e não estruturados). Já Grimes (2008) afirma que a maioria dos dados não estruturados são apenas “não modelados”. Apesar das muitas possibilidades de definição, este trabalho seguirá a convenção geral, considerando como estruturado a forma de tratamento de dados tradicionalmente incorporadas a sistemas de gestão de bancos de dados relacionais. Por consequência, serão considerados como “não estruturados” os dados que ainda apresentam dificuldade de tratamento por esses sistemas, tais como textos, imagens e similares.

Para além da definição formal sobre o que é (ou não) a estrutura de um dado (ou o dado estruturado em si), outra questão de grande importância é se esses dados de estrutura mais complexa de fato possuem valor real para as organizações. As opiniões se dividem. Pessoas contrárias argumentam que o nível de esforço para obtenção de alguma informação é muito grande, e as avaliações ainda estão sujeitas a más interpretações, em função da complexidade

de modelar e da imaturidade de modelos, ainda na atualidade. Já as opiniões a favor justificam que o grande volume de dados disponível compensa o esforço, mesmo que seja para o tratamento de uma fração relativamente pequena desses dados. Estima-se que atualmente sejam criados 2.8 zettabytes (ZB) de dados digitais no mundo por ano e que apenas 1% deles é analisado (Burn-Murdoch, 2012). O IDC divulga uma estatística similar, apontando que dos 40 exabytes existentes hoje, apenas 0,5% é analisado (IDC, 2012). Adicionalmente, o IDC sugere que 80% do crescimento de dados seja realizado por dados não estruturados, sendo eles atualmente responsáveis pela maior participação, conforme apresentado na Figura 3.

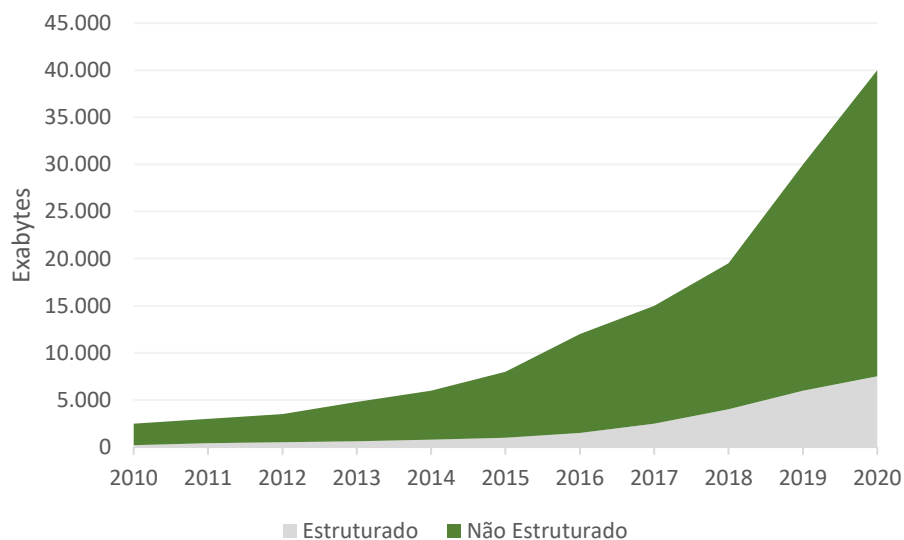


Figura 3: Evolução do volume de dados estruturados e não-estruturados
Fonte: IDC Digital Universe Study (2012)

Naturalmente, o volume de dados não estruturados é muito superior aos estruturados. Um dos fatores que contribui para a naturalidade dessa conclusão é a própria dificuldade de definição do termo “não estruturado”. Qualquer dado, como vídeos e imagens, está incorporado nesse montante, e a modelagem de algoritmos capazes de tratar essas informações representa um desafio contemporâneo. Desse total, é conveniente destacar uma fração desses dados, tipificado como “texto”. Estes são os de maior alvo de esforços para realização de análise, atualmente. Apesar de existirem reconhecidos avanços na modelagem de algoritmos para tratamento destes dados, é possível perceber a insipiência deste tema ainda na atualidade. Chen (2012) faz uma observação específica ao estudo do Quadrante Mágico do Gartner Group, sinalizando que nenhum dos fornecedores destacava, na época, competências de suas ferramentas no tratamento dessas informações. Atualmente alguns avanços são notados no

portfólio de alguns fornecedores (como IBM/Watson/BlueMix e SAP), mas com sensíveis espaços para melhoria. Outros avanços podem ser notados em linguagens como Python e R, que têm sido incorporadas às soluções de fornecedores como forma de fazer suas plataformas evoluírem mais rapidamente.

Um dos importantes aspectos relacionados ao tratamento de dados não estruturados diz respeito ao termo NoSQL (Not only SQL). Este pode ser entendido como uma corrente de pensamento dedicada a pensar em alternativas de aplicações no armazenamento e acesso aos dados, para além do tradicional SQL, que se consolidou na década de 80. Strauch (2011), aponta que a primeira aparição do termo foi em 1998, com o objetivo de se referir a um banco de dados relacional que omitia o uso de SQL. Já em 2009 o termo foi resgatado pelo desenvolvedor Jon Oskarsson para incentivar o uso de bases de dados não relacionais. Posteriormente, Eric Evans ganhou referência, por tornar o termo popular, descrevendo a ambição de um “movimento NoSQL” como uma “busca por alternativas para resolução de problemas, os quais as bases de dados relacionais não foram feitas para resolver”.

Strauch (2011) destaca também que algumas características de bancos de dados NoSQL favorecem a escolha dessa estratégia, citando como exemplo startups e projetos independentes, que evitam os custos de ferramentas proprietárias, e ao mesmo tempo, precisam de flexibilidade e ganho de escala. Ferramentas como Dynamo (Amazon) e BigTable (Google) têm ganhado popularidade nas *startups* tecnológicas, enquanto outros projetos convertem-se para ferramentas de código aberto. Cassandra, que originalmente foi desenvolvido pelo Facebook, hoje faz parte do Projeto Apache.

Como principais benefícios de arquiteturas NoSQL (quando comparadas a bancos de dados relacionais), o autor destaca a possibilidade de evitar complexidades desnecessárias, alto rendimento/desempenho do banco de dados (principalmente com grandes volumes de dados), escalabilidade e performance em hardwares simples (commodities), entre muitos outros aspectos. Os bancos de dados NoSQL possuem diversas formas de implementação, podendo ser utilizadas diferentes taxonomias, de acordo com o modelo de dados utilizado. *Key-Value-Cache*, *Data-Structures Server*, *Document Store* e *Wide Columnar Store* são algumas dessas taxonomias.

Apesar do “movimento” NoSQL ser recente, o potencial dessas iniciativas tem despertado a atenção de alguns autores, inclusive para a formação de profissionais capazes de dominar e operar ferramentas enredadas nestes padrões. Fowler, Godin e Geddy (2016)

discutem a necessidade de inserção de conhecimentos relacionados a NoSQL na formação tradicional de bases de dados e sistemas de informação, principalmente em função das crescentes demandas das organizações no tratamento do tema *big data*.

Um dos fatores que desfavorecem projetos baseados em NoSQL é a falta de padronização de linguagens que manipulem os dados. A evolução dos bancos de dados relacionais foi acompanhada pelo amadurecimento e padronização da linguagem SQL. Contudo, bancos de dados não relacionais ganharam diferentes abordagens e não se consolidou um padrão para linguagens de consultas aos dados. Damien Katz, criador do CouchDB, e uma das ferramentas de bancos NoSQL, e Richard Hipp, criador do SQLite, tentaram formalizar uma linguagem para consulta a bancos NoSQL, batizada como UnQL (Unstructured Query Language). Contudo, esta é ainda uma criação recente, e não necessariamente se consolidará como um padrão de mercado (Bruun, 2012). Outras linguagens para manipulação de dados não estruturados que podem ser citadas são XPATH (para manipular dados XML e eventualmente HTMLs) (Olmedilla, Martínez-Torres, & Toral, 2016) e suas variações (OXPATH, FecthXML) e Earlang. Algumas soluções de bases NoSQL possuem APIs, com linguagens de programação mais tradicionais como Java Script e Python, por exemplo.

Outro aspecto importante é que algumas soluções de NoSQL também suportam a linguagem SQL para a realização de consultas. Como o próprio nome sugere (Not Only SQL – Não apenas SQL em tradução livre), o conceito fundamental por trás dessas iniciativas é a busca por soluções que superem as limitações do SQL, sem necessariamente desconsiderá-lo. Adicionalmente, conceitos/nomenclaturas complementares como NewSQL e ScalableSQL podem ser agregados a esta discussão, de modo a complementar o entendimento sobre as necessidades de evolução do SQL.

2.3.2 Formas de acesso aos dados

A complexidade de questões de negócio colocadas à gestão das organizações sinaliza que o tradicional uso de dados internos não é suficiente para endereçar todos os desafios colocados pela dinâmica do mercado (Ferrández, et al., 2016). Em busca de maior variedade de dados disponíveis para apoio decisório, diferentes iniciativas podem ser buscadas para o

gerenciamento interno dos dados oriundos de fontes externas. Alguns exemplos são projetos de diversificação de Data Warehouse como “*Exploratory OLAP*” (Ibragimov, Hose, Pedersen, & Zimanyi, 2015), ou modelos de arquitetura de dados alternativos, como *Data Federation* (Nabi, Sabir, Bilal, & Ayub, 2017). As recentes necessidades pela diversificação de dados contribuíram também para a intensificação do campo de estudo de Governança de dados, inclusive no contexto de projetos de BI (Teodoro, Przybilovicz, & Cunha, 2014). Apesar dessa diversificação, iniciativas dessa natureza tem encontrado grandes dificuldades de implementação. Assim, destaca-se como um grande desafio o acesso aos dados e a conjugação de diferentes fontes de dados, internos e externos à organização (Randall & Beyer, 2015).

Anderson (2008) argumenta que a melhor acessibilidade dos dados está fomentando um ambiente de interesse no tema de *big data*. Apesar dos notáveis avanços de volume de dados existentes atualmente, e do consequente aumento de disponibilidade dos dados aos indivíduos, cabe uma reflexão mais profunda sobre a real acessibilidade dos dados, sobretudo daqueles pertinentes aos negócios. De forma complementar ao debate das formas de acesso, é necessário que se tenha também uma perspectiva sobre o tema de compartilhamento de dados. Wells (2016) defende que o compartilhamento de dados pessoais é uma questão-chave para que sejam criados serviços que beneficiem a nós e à sociedade. Por outro lado, dados pessoais não são o único aspecto a nos atentarmos. Empresas como Amazon, Facebook, além de startups como Spotify, Airbnb e Foursquare estão aderindo ao compartilhamento de suas informações. Kelly (2014) argumenta que existe um mercado de dados se formando, e que as empresas tenderão a atuar neste mercado, dinamizando as informações de suas indústrias/setores, com o objetivo de obter produtos e serviços com cada vez maior qualidade. A proposição de Redman (2008) reforça esse argumento, sugerindo que uma forma de os dados da organização provarem seu valor é expondo-os ao mercado.

O desafio de acessar diferentes fontes de dados pode ser observado sob o espectro de compartilhamento dos dados. Neste, diferentes níveis de acesso devem ser considerados, construindo, assim, uma referência para a caracterização das fontes de dados. O Instituto de Dados Abertos (*Open Data Institute* – ODI) apresenta um modelo de referência especificamente orientado à discussão do compartilhamento de dados. No modelo é considerado o nível mais restritivo como o nível de “dados fechados”, passando por um estágio intermediário de “dados compartilhados”, e alcançando o nível mais permissivo, classificado como “dados abertos”, tal como evidencia a Figura 4 a seguir:

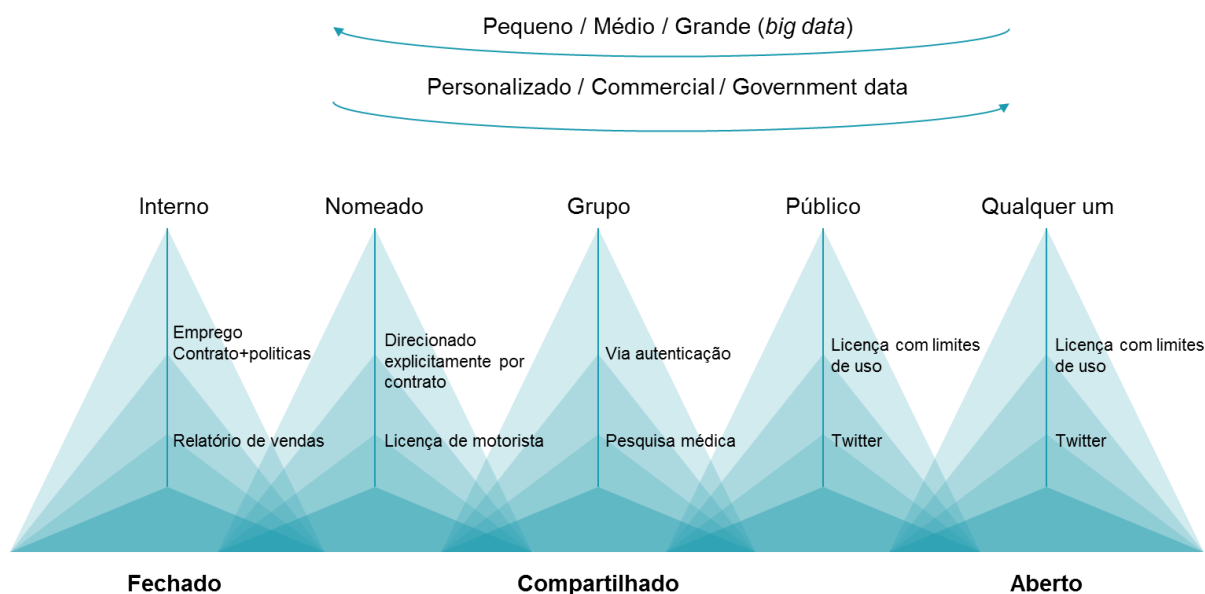


Figura 4: Espectro de compartilhamento de dados

Fonte: *Open Data Institute - The Data Spectrum*, traduzido pelo autor

De forma complementar ao proposto pelo ODI, um conjunto de autores discutem algumas referências de tipos de fontes de dados disponíveis para acesso por parte das organizações. Apesar de não ter encontrado uma referência na literatura acadêmica que construa uma tipificação explícita das fontes de dados, o conjunto de referências sugere a organização das fontes de dados conforme estruturado na Tabela 1.

Escolher quais fontes de dados estarão incorporadas ao projeto de IA pode variar em função das iniciativas adotadas, bem como pelos resultados desejados. Deste modo, é necessário que a empresa tenha claro que, para cada uma dessas classificações, existem características específicas pertinentes a ela. Por exemplo, dados governamentais são muito permissivos em termos de custos, mas pode apresentar estruturas de dados complexas, ou uma estrutura de suporte frágil (como a ausência de metadados), por exemplo.

Nesse sentido, é importante que se analise as diferentes fontes de dados que estão disponíveis para uma iniciativa de IA. Isso deverá auxiliar a organização na identificação de *trade-offs* para a escolha de fontes de dados e estratégias a serem seguidas. Se por um lado os dados abertos do *Governo* podem apresentar alta complexidade em sua estrutura, alternativamente esses dados podem ser transacionados através de intermediadores a custos relativamente atrativos. Experiência similar a esta pode ser verificada pela empresa Socrata, que atua nos Estados Unidos como intermediária de informações do governo (Socrata, 2017).

Tabela 1: Tipos de fontes de dados

Tipo de fonte de dados	Descrição da fonte de dados	Literatura relacionada
Dados Internos	São produzidos em mantidos restritos ao ambiente interno das empresas	Chen, Chiang, & Storey (2012); Abbasi, Sarker, & Chiang (2016); Willcocks, Venters, & Whitley (2014)
Dados compartilhados de outros negócios	Empresas que compartilham parte de seus dados, sem que essa seja seu “ <i>core business</i> ”, ou uma importante fonte de receita	Redman (2008); Chen, Chiang, & Storey (2012); Olmedilla, Martínez-Torres, & Toral (2016); Boyd & Crawford (2012); Albertin (2001)
Intermediadores/Facilitadores	Agentes que fazem intermediação de informações no mercado como parte do seu “ <i>core business</i> ”	Redman (2008); Balazinska, Howe & Suciu (2011); Faria, Linden, & Laney (2017); Natis (2016)
Dados oriundos do Governo	Dados coletados e distribuídos pelo governo. Usualmente são dados abertos.	Jetzek, Avital e Bjørn-Andersen (2013); Hofman & Rajagopal (2014); Sadiq & Indulska (2017)

Apesar dos benefícios potenciais da utilização de dados externos, os autores Abbasi, Sarker e Chiang (2016) destacam fragilidades no seu uso, sendo a principal delas é que a credibilidade das diferentes fontes varia. Por exemplo, dados oriundos de mídias sociais são comumente contaminados por *spam*, enquanto *web spam* chega a contabilizar mais de 20% de todo o conteúdo na *World Wide Web* e 40% de todas as páginas no domínio *.US* e *.Biz* não são confiáveis (Abbasi & Adjero, 2014). *Clickstreams* de navegação em *web sites* e *mobile* são muito suscetíveis a ruídos. Adicionalmente, os resultados de Isik, Jones e Sidorova (2012) confirmam que usuários de Inteligência Analítica são menos satisfeitos com dados externos. Nesse estudo são considerados critérios de confiabilidade, consistência e precisão, comparando dados externos com internos. Eles destacam que menos de 40% da amostra trabalhada expressaram satisfação nesses critérios.

Nos itens a seguir são apresentadas algumas características relacionadas aos tipos de fontes de dados, conforme citado na Tabela 1, com o objetivo de enriquecer o entendimento acerca de cada um desses elementos.

- ***Dados Internos***

Os dados limitados ao contexto da empresa, sem a utilização de terceiros, são considerados dados internos. Este tende a ser o tipo mais tradicional de fontes de dados em projetos de Inteligência Analítica, e amplamente utilizado nas organizações (Chen, Chiang, & Storey, 2012). Trata-se, portanto, de uma estrutura tradicional de TI para armazenamento e consumo de dados, com tráfego dentro do domínio da organização.

Apesar de os dados serem de propriedade da empresa, esta classe passa hoje por mudanças que valem ser mencionadas. A primeira delas é o fortalecimento das migrações de infraestrutura para ambientes remotos, conhecidos como *Cloud Computing* (Willcocks, Venters, & Whitley, 2014). Assim, apesar de serem proprietárias dos dados, por vezes as empresas poderão enfrentar limitações de ordem técnica, ou burocrática, na relação com o seu fornecedor de soluções no ambiente *Cloud*. A segunda questão a se destacar é o crescente consumo de *Software as a Service (SaaS)*, movimento que passa a hospedar os dados das empresas necessariamente numa infraestrutura externa, sendo o acesso aos dados limitados às condições do provedor dessa solução. Em geral, integrações de sistemas via *Application Programming Interface – API* são necessárias.

O maior potencial desse tipo de fonte a ser explorado, presente na literatura, são os dados não estruturados. Durante muito tempo os sistemas transacionais nas organizações acumularam dados, sobretudo textuais, referentes a diferentes interações com o cliente. Além de processos comerciais, as interações de atendimento e pós-venda registram informações importantes sobre os clientes, e também sobre os próprios processos. Assim, muitas organizações possuem esse real potencial de análise dos dados, podendo conhecer ainda mais a fundo o próprio negócio. Apesar de serem muitas as possibilidades, é necessário admitir ainda que o entendimento semântico de textos continua sendo um desafio em muitas situações, apesar dos significativos avanços em Processamento de Linguagem Natural (Abbasi, Sarker, & Chiang, 2016).

- ***Dados compartilhados de outros negócios***

Na categoria de dados compartilhados de outros negócios, as informações transacionadas são proprietárias, mas compartilhadas com terceiros. O modelo de compartilhamento pode prever políticas específicas de acesso, como restrição a usuários específicos, custos para consumo dos dados, ou o uso de uma tecnologia em particular. A principal característica desta categoria é que o compartilhamento de dados não faz parte central de seu modelo de negócio. as empresas compartilham suas informações em função de outros benefícios vinculados à sua operação, que não visam necessariamente à compensação financeira direta. Assim, o compartilhamento de seus dados assume o papel de agregador de valor, junto aos produtos e/ou serviços oferecidos pela organização (Redman, 2008).

Diferentes razões motivam uma empresa a disponibilizar suas informações. Dependendo do setor de atuação, ou de determinadas políticas adotadas (como financiamento pelo mercado de capitais, por exemplo), a empresa pode ser motivada a compartilhar parte de seus dados em função de requisitos legais ou normativos. Dependendo da legislação e da empresa, a divulgação de informações contábeis pode também ser um requisito legal. Contudo, motivações não vinculadas a obrigações legais e/ou regulatórias existem, como o aumento da transparência junto a seus investidores e/ou sociedade, trocas de informação dentro da indústria para fortalecimento de sua cadeia de suprimentos, divulgação de produtos e serviços, atuação no comércio eletrônico (Albertin, 2001), entre outros.

No início do desenvolvimento da infraestrutura Web poderia ser difícil imaginar uma organização motivada a compartilhar seus dados de forma espontânea. Um exemplo é o serviço *Amazon Web Services*, que permite que desenvolvedores acessem seu catálogo de produtos, avaliações de consumidores, pontuações do site e preços históricos (Olmedilla, Martínez-Torres, & Toral, 2016). Ao permitir acesso aos dados, a empresa possibilita que integradores possam comparar suas ofertas com outros players, ampliando seu alcance de mercado.

Outro exemplo de compartilhamento de dados são as Mídias Sociais. Empresas como Facebook e Twitter podem perceber, no compartilhamento de suas informações, uma maior capacidade de integração com plataformas corporativas e um consequente aumento de tráfego ou capacidade de monetização de seus serviços. Outras empresas podem tratar seus dados como produto final, de forma direta. Dados cadastrais de pessoas físicas ou jurídicas, por exemplo, podem ser utilizados para gerar inteligência de marketing aos compradores dos dados, como é o caso da empresa Boa Vista, com as Associações Comerciais no mercado brasileiro.

O caso recente do Pão de Açúcar é emblemático para caracterizar as novas possibilidades de compartilhamento de dados. A empresa adotou como estratégia de negócio trabalhar a fidelização do cliente, buscando oferecer ofertas mais aderentes aos desejos do consumidor. A novidade está na forma como a empresa implementou esta estratégia. Os dados dos clientes cadastrados no programa de fidelidade foram repassados aos fornecedores, permitindo que a rede de supermercados apresentasse ofertas de interesse de seu público, financiadas por seus fornecedores.

Tendo acesso ao perfil de quem consome seus produtos, os fornecedores podem fazer ofertas direcionadas. Ao falar sobre a iniciativa, o gerente do programa de fidelidade afirma que “a indústria consegue entrar (na plataforma criada) e verificar o cliente que ela quer impactar: os mais fiéis ou menos fiéis, os abandonadores...”. A ação permite que os fornecedores consigam fazer iniciativas direcionadas ao público que querem impactar. O responsável pela iniciativa exemplifica, mencionando que o fornecedor “pode fazer um desconto só para quem compra a categoria, mas não compra seu produto [...] ou para quem compra o produto de menor valor da marca e poderia subir para uma categoria *premium*”. Indagado sobre o que acontecerá quando os concorrentes imitarem a estratégia, o entrevistado se diz confortável, por estar na liderança, e afirma que essa “é a evolução do varejo: no mundo da informação não dá mais para ter aquele ‘*black box*’ com o fornecedor, como era nos anos 80” (Viri, 2017).

Poucos meses depois, o Instituto Brasileiro de Defesa do Consumidor (Idec) notificou o grupo Pão de Açúcar, em função da ausência de políticas de privacidade nos aplicativos utilizados para coleta dos dados dos clientes. Para ter acesso a preços promocionais, o consumidor precisa baixar um aplicativo em seu celular e fazer um cadastro com informações como gênero, endereço e número de filhos. O questionamento feito pelo Idec considera que o aplicativo não oferece informações claras, relacionadas a políticas de privacidade, ferindo o Marco Civil da Internet. Além disso, indicam que o aplicativo solicita acesso a informações desnecessárias para seu funcionamento, como fotos e arquivos armazenados no dispositivo, geolocalização e câmera (Instituto Brasileiro de Defesa do Consumidor, 2017).

Apesar de o compartilhamento de informações poder ser algo positivo para profissionais do mercado de Inteligência Analítica, é importante que se tenha atenção aos dados consumidos. Empresas possuem a liberdade de compartilhar as informações, de acordo com a conveniência, especialmente se estiverem livres de obrigatoriedade legal. Deste modo, é difícil que se tenha

algum controle sobre a qualidade de dados que são consumidos. Além da omissão de registros, outras dificuldades técnicas, como ausência de metadados ou documentação técnica, pode dificultar a interpretação das informações.

Boyd e Crawford (2012) argumentam que, no âmbito da pesquisa acadêmica, o pesquisador deve atentar ao fato de que é difícil compreender a amostra quando a fonte de dados é incerta. Eles afirmam que o Twitter, por exemplo, disponibiliza apenas uma fração de seu material para o público, por meio de suas APIs. Teoricamente, o fluxo de dados obtidos pela API deveria conter todos os registros considerados como “públicos”, protegendo apenas aqueles conteúdos classificados como “privados” pelos usuários. Ainda assim, afirmam que nem todos os conteúdos classificados como “públicos” são capturados pela API (Boyd & Crawford, 2012).

- ***Intermediadores/Facilitadores***

Algumas entidades atuam no mercado como intermediadores de dados. Trata-se de um mercado fragmentado, podendo atender a diferentes nomenclaturas, como *data brokers* ou *data hubs*. Seu modelo de negócio não tem como objetivo final divulgar seus próprios dados, mas se dedica, principalmente, a agregar e distribuir conteúdo já disponível no mercado.

As formas mais usuais de agregar valor são: facilitação do consumo dos dados (estrutura ou processo de coleta) e/ou; adição de algum nível de análise (como criação de índices, por exemplo). Os modelos de negócio podem buscar aplicar custos aos seus clientes, ou trabalhar de forma híbrida (cobrando apenas um determinado conjunto de dados, ou pelo volume de uso, por exemplo). Idealmente, atores nessa classificação devem buscar uma estrutura de dados suficientemente simples, com ferramentas de consumo dos dados adequadas ao seu público alvo. Adicionalmente, um fornecedor que se apresente como intermediador deve buscar um modelo de precificação adequado para seu negócio, entregando valor ao cliente de modo a viabilizar o negócio. Assim, deve apresentar como uma solução de negócio vendável. Adicionalmente, deve facilitar o processo de coleta das informações (Balazinska, Howe, & Suciu, 2011).

Apesar de a precificação ser um aspecto importante na decisão de transacionar dados, as fontes classificadas como *Intermediadores* são os que devem estar mais atentos a esta questão, uma vez que a transação dos dados é a essência de seu negócio. Balazinska, Howe e Suciu (2011) apontam a formação de um mercado de dados como uma oportunidade para a comunidade de Base de Dados, na qual dados em nuvem viabilizam o compartilhamento de

informações e podem ajudar os usuários a encontrar e consumir dados. O trabalho destaca também a importância de modelos de precificação para esse mercado, em que dados brutos podem ser precificados de forma diferente de “dados derivados”, além da escolha (ou não) da variação de preços em função do volume acessado.

O portal da Amazon (*Amazon Web Services*) agrega um conjunto de fornecedores de dados abertos, no formato de parceria. Os dados transacionados são oferecidos pelos parceiros, que utilizam a plataforma da Amazon como infraestrutura intermediária. Plataformas como essa têm surgido no mercado com o objetivo de agregar múltiplas fontes de dados e oferecê-los como serviço ao consumidor final. Alguns fornecedores de ferramentas de *Business Intelligence*, como Qlik e Microsoft, apresentam soluções similares a essa (*Qlik Data Market* e *Windows Azure Datamarket*, respectivamente).

No contexto mais recente surgiram trabalhos sobre plataformas de compartilhamento de dados, como serviços (*Data Broker Platform as a Service*) na plataforma do Gartner, na qual o tópico passou a figurar o estágio inicial na análise “*Hype Cycle*” para plataformas como serviços, oferecida pela instituição. Essas plataformas têm como objetivo desenvolver um ambiente onde diferentes atores possam disponibilizar seus dados, oferecendo ao consumidor final a capacidade de suportar processos de Inteligência Analítica (Natis, 2016). Outro trabalho publicado pelo grupo descreve a necessidade de que as empresas reflitam sobre a adoção de fornecedores de dados. Afirmam que, apesar das empresas precisarem de fontes externas em maior número e qualidade, esse mercado é muito fragmentado, em alto crescimento e com mudanças constantes (Faria, Linden, & Laney, 2017). Algo que adiciona complexidade a este mercado é que cada fornecedor possui uma combinação única de fontes de dados exploradas, modelos de precificação, termos de licenciamento, formatos de dados e serviços relacionados, dificultando a comparação entre os fornecedores. O mesmo estudo considera que, num futuro próximo (2020), as agências do *Governo* tenderão a gerar receitas por meio de dados com valor agregado, transacionados nesse mercado.

Outros players não vinculados a ferramentas de Business Intelligence podem ser citados, como QuandL, IndexMundi e World Bank. O QuandL iniciou suas atividades em 2013 apenas como um concentrador (Data Hub) de informações públicas diversas. Hoje, opera como uma plataforma que, além de disponibilizar dados públicos, também permite a comercialização destas informações (QuandL, 2017). Diferentemente das outras plataformas mencionadas, ela disponibiliza uma documentação robusta, livre de custos para o usuário. A monetização da

plataforma acontece apenas para fontes classificadas como “não gratuitas”. Para usuários “não desenvolvedores”, oferece uma integração com o Excel, capaz de automatizar a obtenção de dados. Para o público de desenvolvedores, disponibiliza APIs para coleta de informações em Python, com bibliotecas especificamente desenvolvidas para essa finalidade, fator que criou engajamento da comunidade de desenvolvedores, no contexto de Big Data. O IndexMundi disponibiliza informações de forma similar, envolvendo dados abertos, como preços de commodities e dados demográficos, por exemplo. Se apresenta em plataforma web, não oferecendo ao usuário a funcionalidade de APIs, tornando o processo de coleta de dados mais complexo (caso o usuário deseje automatizar a coleta de dados, terá de trabalhar com instrumentos de *data crawling*). Já o WorldBank apresenta indicadores globais em diferentes temas e recortes geográficos, por meio de sua plataforma web, ofertando também a possibilidade de automação, através de APIs, tudo de forma gratuita.

Alguns fornecedores trabalham de forma híbrida, combinando dados brutos com “informações derivadas” (adicionando dados produzidos pela própria empresa e/ou análises derivadas de dados brutos). Num contexto internacional, fornecedores como Bloomberg e S&P/CapitalIQ são mais conhecidos. No Brasil podemos mencionar o ValorPRO e Econodata, como exemplos de atores. Especializadas no mercado financeiro, estas plataformas oferecem um serviço de dados que facilita a navegação do usuário e a automação da coleta de dados. Os dois primeiros casos possuem integração com aplicações do Office (Microsoft), além de também disponibilizar APIs. Os dois atores brasileiros mencionados oferecem o consumo dos dados de forma limitada ao uso de uma plataforma proprietária, apenas com a funcionalidade de exportação de dados de forma manual.

- **Governo**

Dados abertos podem ser definidos como aqueles disponíveis para serem livremente utilizados, reutilizados e redistribuídos por qualquer um e submetidos apenas (quando muito) à obrigação de atribuir citação a fonte dos dados (OKF, 2012). O debate envolvendo o governo é particularmente importante, em função de sua centralidade, volume de dados que armazena e pelo papel intrínseco que exerce na sociedade de estimular o desenvolvimento econômico e social.

Dados abertos disponibilizados pelo governo podem oferecer valor econômico para a iniciativa privada e devem ser considerados como uma alternativa viável a ser trabalhada pelas organizações. Jetzek, Avital e Bjørn-Andersen (2013) argumentam que dados abertos oriundos

do governo são um subconjunto de dados particularmente interessante, uma vez que já foram coletados para um uso específico, foram pagos pelos contribuintes, são relevantes e oferecem valor para além do objetivo que foram coletados originalmente. Quando são abertos, se tornam um recurso compartilhado e comum, fornecido pelo governo.

Assim, dados abertos possuem as seguintes características: 1) Livre de concorrência; 2) São “não exclusivos”; 3) Possuem um alto custo fixo; 4) Possuem um custo marginal (quase) zero; 5) Oferecem informação valiosa e; 6) possuem alto potencial de reutilização. Também apontam que um conjunto de elementos por parte do governo podem oferecer barreiras para a criação de valor a partir dos dados: 1) Conjuntos de dados fechados ou inacessíveis/restritos; 2) Falta de uma política de dados robusta; 3) Falta de validade, completude ou exaustividade dos dados; 4) Insuficiência de metadados, bem como falta de interoperabilidade técnica ou semântica; 5) Falta de consistência no acesso para além de fronteiras (cross-border); 6) Falta de motivação do setor público; 7) Falta de competência técnica no setor público; 8) Falta de “alfabetização de dados” e habilidade técnica; 9) Comunidade de dados abertos muito fragmentada e separada (Jetzek, Avital, & Bjørn-Andersen, 2013).

A interoperabilidade técnica e semântica constitui-se como um desafio particular para maior aproveitamento de dados públicos governamentais pelas empresas. Cada provedor de dados é autônomo e tem uma implementação particular de processos e solução de TI, apesar de muitas instâncias (como as municipais) gerenciarem dados de natureza similar e que poderiam utilizar soluções idênticas. As diferenças de sintaxe e semântica podem decorrer em função dessa diversidade de projetos, ou de definições implícitas (como atributos de documentos XML) com as quais o usuário deve lidar (Hofman & Rajagopal, 2014).

Em 2011, o Laboratório Brasileiro de Cultura Digital e o Núcleo de Informação e Coordenação do Ponto BR – NGI-BR publicaram uma versão do Open Data Manual, adequando-o para o contexto brasileiro. No documento destacam a escassez de instituições públicas que trabalham com dados abertos e a necessidade de ampliar esse debate no país, apesar dos reconhecidos avanços, como a Lei de Acesso à Informação, por exemplo. Alguns exemplos nacionais do uso de dados públicos governamentais envolvem assuntos como transparência orçamentária ou mecanismos de investigação de interesse social, como pontos de alagamento na cidade ou controle de reclamações junto à prefeitura. Nos casos apresentados destaca-se o esforço necessário dos desenvolvedores na obtenção dos dados, que estão disponíveis, porém através mecanismos que dificultam seu acesso. Nesse sentido, reflexões

sobre características dos dados podem auxiliar na definição técnica de mecanismos para acesso aos dados, como APIs, *web services* e bancos de dados (Comunidade Transparência Hacker, 2011).

Existe atualmente um forte debate sobre a importância de se atentar à qualidade dos dados abertos disponibilizados pelo governo. Um dos movimentos existentes é a definição de critérios consistentes capazes de classificar o quão “aberto” são os dados, e que nível de qualidade eles possuem. Desse modo, atenta-se a granularidade dos dados, sua representatividade em diferentes recortes territoriais, quantos segmentos de interesse público o governo é capaz de atender (saúde, educação etc.), entre outros aspectos. Nesse sentido, discute-se a criação de um índice de qualidade dos dados abertos, para estimular que diferentes governos sejam capazes de perseguir uma estratégia de dados para sua comunidade (Sadiq & Indulska, 2017). Outro aspecto a ser discutido são as características técnicas de compartilhamento de dados pelo governo. Modelos de referência devem ajudar governos de diferentes esferas, bem como entidades de diferentes tipos, a optar por tecnologias mais compatíveis entre si, minimizando custos de transação entre as diferentes fontes de dados (Hofman & Rajagopal, 2014).

2.3.3 Métodos de obtenção de dados

No passado mais recente a tecnologia aumentou sua influência sobre os métodos de coletas de dados. A consolidação da internet, assim como sua ampla utilização por parte da sociedade, possibilita que hoje muitas coletas de dados sejam feitas por meio de computadores. Atualmente, discute-se a possibilidade de coletas de dados na internet integrarem parte de pesquisas de preço (tal como para a inflação) conduzidas por órgãos do governo, por exemplo (Polidoro, Giannini, Conte, Mosca, & Rossetti, 2015).

Assim, esta sessão se dedicará a avaliar dois métodos específicos de obtenção de dados disponíveis na internet. São provavelmente os mais importantes métodos relacionados ao tema, que devem ser objeto de atenção e domínio de profissionais que trabalham na área. O primeiro, utilizando APIs e *web services*, parte da premissa que o fornecedor de dados construiu uma aplicação, de tal modo que existam funções e padrões de conexão pré-definidos para facilitar o

acesso às informações. O segundo, chamado de *Web Scraping* (entre outras nomenclaturas), utiliza códigos de programação para simular a navegação de um usuário comum com o objetivo de acessar os dados contidos num determinado site.

Estes tipos de coletas de dados são utilizados na indústria e começam a ser incorporados em metodologias de trabalhos acadêmicos, em função do grande volume de dados a que se tem acesso (Olmedilla, Martínez-Torres, & Toral, 2016). Apesar disso, existem ressalvas importantes a serem consideradas no contexto acadêmico, dado que determinadas fontes levantam dúvidas sobre a real consistência dos dados disponíveis nelas.

De acordo com Olmedilla, Martínez-Torres e Toral (2016), devem ser feitos destaques para justificar a importância da utilização desses métodos como instrumentos de coletas de dados pelas organizações, a saber:

- Reduz drasticamente os custos de coleta de informação das empresas, pois obtém dados que já foram preenchidos por alguém
- Amplia consideravelmente o tamanho da amostra
- Traz maior poder de análise, pois possibilita a agregação de diferentes fontes
- Como limitação, está restrito aos dados que já foram preenchidos por um usuário, num contexto externo a uma pesquisa direcionada (não existe o planejamento de um questionário, por exemplo)

APIs e Webservices

Um dos aspectos que fortalece a construção de um mercado de transação de dados é a facilitação do processo de coleta de dados. No contexto da Web 2.0, um conjunto de conceitos auxilia e fortalece esse processo, na medida em que facilita a integração de sistemas e consumo de dados. *Web Services* baseados em protocolos de internet (HTTP, SMTP e XML), além de APIs associadas a estes serviços, permitem que desenvolvedores integrem diversos conteúdos da Web (Chen, Chiang, & Storey, 2012). Essas integrações possibilitam novas formas de uso dos dados, sobretudo no contexto da Inteligência Analítica.

No conjunto de conceitos que envolve a disseminação de dados em serviços de internet, tais como AJAX, RSS e REST, as APIs (*Application Programming Interface*, ou Interface de Programação de Aplicações) apresentam um papel central, nesse contexto. Uma API é criada quando um serviço tem a intenção de que criadores de software desenvolvam produtos associados a esse serviço. Vários deles disponibilizam seus códigos e instruções para serem

usados em outros sites, da maneira mais conveniente para seus usuários. O Google Maps é um dos grandes exemplos na área de APIs. Por meio de seu código original, muitos outros sites e aplicações utilizam os dados do Google Maps, adaptando-o da melhor forma a fim de utilizar esse serviço (CanalTech, 2017).

O conceito de APIs está orientado para a integração de serviços, não necessariamente ao processo de coleta de dados. Contudo, essas integrações são feitas fundamentalmente mediante a troca de dados, o que faz com que esta ferramenta se torne adequada para a troca de informações entre diferentes sistemas. Apesar de ser um conceito relativamente antigo, os avanços da computação em nuvem e das aplicações móveis (*mobile*) favoreceram a popularização do uso de APIs (Taulli, 2015). O aumento do número de APIs disponíveis enriquece o ambiente de transação de dados e permite que negócios migrem para plataformas digitais, oferecendo produtos e serviços a um número maior de clientes, eventualmente combinando-se a outros produtos e serviços desenvolvidos por outro negócio. Essas novas possibilidades oferecidas ao negócio são objeto de atenção de líderes de tecnologia nas organizações (Malinverno, Moyer, O'Neill, & Gilpin, 2017). Contudo, a forma como os negócios podem ganhar dinheiro com as APIs não está diretamente relacionado à tecnologia em si. Em outras palavras, as receitas geradas por um negócio que utiliza APIs não estão vinculadas ao uso da API propriamente, mas às oportunidades de negócio possibilitadas por elas (Malinverno, 2016). Sendo assim, a funcionalidade das APIs tende a se consolidar como fundamental no desenvolvimento de softwares, dado que elas podem assumir um valor central no modelo de negócio de algumas organizações: Salesforce gera 50% de suas receitas através de APIs; Ebay aproximadamente 60% e Expedia 90% (Murphy, 2016).

Apesar da facilidade da coleta de dados promovida pelo uso das APIs, há quem apresente ressalvas sobre o uso dessa tecnologia como instrumento de coleta, sobretudo para condução de pesquisas acadêmicas. Olmedilla et al. (2016) destacam que, embora APIs facilitem a extração de conteúdos de forma automática, elas possuem algumas limitações quando se deseja acessar dados específicos requisitados por pesquisadores. Ocorre que as APIs apenas facilitam a obtenção das informações que o provedor de conteúdo decide disponibilizar. Em seu estudo, Olmedilla et al. (2016) buscaram métodos alternativos de coleta de dados para superar as limitações das APIs, conforme apresentado na seção a seguir.

Para além das APIs: o papel do web scraping

Se, por um lado, a popularização de APIs é um importante avanço nas possibilidades de acesso às informações transacionadas pela internet, é necessário que às organizações estejam atentas às possibilidades de avançar para além deste método. A motivação para ir além se deve a, basicamente, dois fatos. O primeiro é que nem todos (na verdade, são poucos) os sites (ou serviços da web) disponibilizam APIs. O segundo é que, mesmo naqueles sites que oferecem o método, por vezes, os dados disponibilizados são apenas uma fração do que pode ser obtido (Olmedilla, Martínez-Torres, & Toral, 2016).

A possibilidade de extração de dados, a partir de conteúdos de websites disponibilizados na internet, é explorada há muito tempo, e ganhou maior popularidade a partir do êxito dos Motores de Busca, ainda no início da internet. Parte importante dos Motores de Busca são os algoritmos chamados de *Web Crawlers*, responsáveis pela indexação e organização de conteúdos disponíveis na Web. O primeiro *Web Crawler* foi criado em 1993, com o objetivo de compilar estatísticas a respeito do crescimento da web (Najork, 2009). Esses algoritmos são definidos como um processo metódico de varredura das páginas disponíveis na web, fazendo o tratamento e indexação das informações encontradas. Além da indexação de conteúdo ser útil, como prestação de serviços para usuários que buscam encontrar conteúdos, as empresas encontraram mecanismos de monetização desse serviço, através de publicidade direcionada de acordo com o conteúdo acessado pelo usuário e pelo perfil deste.

Além dos *Web Crawlers*, existem outros termos que definem algoritmos de natureza similar. Por exemplo, uma atividade que ganhou a atenção de desenvolvedores nas últimas décadas é o *Web Scrapping*. De forma análoga, esses algoritmos são definidos por um processo sistemático de captura de informações, a partir de um conjunto de URLs denominadas como *seeds* (sementes). A partir dessas URLs, o código é capaz de extrair os conteúdos desejados, além de navegar para novas URLs, localizadas a partir das sementes, obtendo todo o conteúdo desejado pelo desenvolvedor ou operador/usuário. Outra variação desse processo é denominada de *Wrapping*, um algoritmo capaz de converter as informações obtidas numa determinada fonte para a forma relacional, dialogando com estruturas de dados mais tradicionais (conforme discutidas em seções anteriores).

Os conceitos de *Web Crawlers*, *Web Scrapping* e *Wrapping* são próximos, e por vezes podem se misturar. Como mencionado por (Najork, 2009), *Web Crawlers* podem ser utilizados para além das funções de indexação de informações, oferecendo serviços que mineram a *web*

em busca de violações de direitos autorais, ou serviços de comparação de preço, entre outros. Esses serviços poderiam também ser entendidos como *Web Scrapping/Wrapping* (dado que eles obtêm informações e oferecem análise a partir delas).

Algumas pesquisas acadêmicas têm se preocupado com a técnica de *web scraping* como método para coleta de dados. Por exemplo, o trabalho Olmedilla, Martínez-Torres e Toral (2016) captou informações de avaliações de produtos gerados por pessoas num website chamado Ciao, classificado como uma grande comunidade de *eletronic Words of Mouth* - eWOM¹. Contudo, a autonomia de coleta de dados através de processos automáticos deve se estabelecer como um ponto de atenção para as organizações. Para além do uso de APIs, as técnicas de *web crawling* permitem que as empresas acessem ainda mais dados sobre o mercado e são capazes de estruturar um monitoramento sistemático de seus concorrentes, parceiros e clientes. Este ponto de vista é defendido por Glez-Pena, *et al.* (2013). Eles destacam que, apesar da técnica de *web scraping* ser relativamente antiga, ela ainda é capaz de desempenhar um importante papel na obtenção de dados, independentemente da disponibilidade de APIs.

O amadurecimento dessa técnica permite que hoje exista um conjunto de ferramentas disponíveis para usuários e programadores, que possibilitam a implementação desses algoritmos em diversas linguagens, como PHP, Perl, Java, Javascript e Python (Chalk, 2015). Em Python é possível realizar processos de *web crawling* usando bibliotecas como *beautifullsoap*, *selenium*, e um *framework* colaborativo e *open source* chamado *Scrapy*, por exemplo.

Os autores Glez-Pena, *et al.* (2013) apresentam uma revisão de literatura, destacando não apenas as bibliotecas e frameworks disponíveis para diferentes linguagens de programação gratuitas, como também o amadurecimento de ferramentas prontas de *web scraping* para usuários “não-programadores”.

Apesar de as técnicas de *web scraping* trazerem maior versatilidade para a obtenção de dados, existem debates sobre a validade ética de práticas como essa. Liu e Davis (2015) expõem

¹ Words-of-Mouth (WOM) é um termo utilizado para descrever a transmissão de informações de pessoa para pessoa, através de comunicação oral. No campo de estudo de Marketing, a expressão é utilizada para descrever a comunicação feita entre comunicadores “não comerciais”. Ao utilizar uma plataforma digital para transmitir a informação, a expressão ganha a variação para Eletronic Words-of-Mouth (eWOM)

a visão de que os dados colocados pelas empresas em seus *web sites* são como “amostras grátis”. Nesse contexto, um visitante que abuse do consumo dessas “amostras” poderia ser considerado como um visitante indesejado. Esse raciocínio é construído no contexto da internet, onde as escalas de práticas nesse sentido podem ser muito maiores. Eles apresentam também, em seu texto, algumas discussões sobre legislações (envolvendo Estados Unidos e Europa), e discutem instrumentos de proteção que podem ser utilizados pelas empresas. Um deles é a declaração de “Termos e Condições” nos sites, com respectiva aceitação por parte do usuário. Contudo, esse instrumento, para ter validade, deve ser aplicado de forma compulsória aos usuários do site, de modo que acessem as informações apenas se concordarem, fazendo com que esse artifício perca atratividade.

Também é possível que se apliquem mecanismos técnicos que dificultem a captura de dados, como instrumentos que bloqueiam o conteúdo e utilizem, como chave para o acesso à informação, o reconhecimento de imagens. Outros mecanismos mais inteligentes podem ser capazes de reconhecer o padrão de navegação do usuário e impedir que aquele acesso continue visitando o conteúdo da página. Contudo, ao passo que as tecnologias para impedir o uso de “robôs” evoluem, os algoritmos de coleta também evoluem, embaralhando padrões de navegação, ou mesmo utilizando reconhecimento de imagem.

Para além da questão ética, as atividades de *web scraping* trazem implicações técnicas para o funcionamento dos sites que são seus alvos. O número de acessos ou a realização de atividades indesejadas na página podem prejudicar o desempenho do servidor ou ocasionar mau funcionamento para outros usuários. Essa é uma questão que acompanha os desafios dos primeiros *web crawlers* e ainda permanece: de que forma posso acessar os dados de forma “educada”, ou seja, minimizando o custo computacional do servidor e otimizando minha operação.

A empresa ScrapeSentry, especializada em soluções para proteção de *web scraping* para conteúdos disponíveis na internet, destaca em seu relatório anual que, em 2014, o tráfego de algoritmos nas páginas da internet aumentou 17%, em relação ao ano anterior (Scrape sentry, 2015). Também destaca que 22% de todo o tráfego pode ser considerado como *scrapers* e que 49% desse tráfego são originados nos Estados Unidos e 13% na Suécia. Ainda segundo este relatório, a principal indústria suscetível à cópia dos dados é a de Turismo/Viagens, seguida por Classificados Online (vendas de automóveis, sites de compra e venda, e outros).

Se, por um lado, ser “vítima” de *web scraping* pode ocasionar dificuldades operacionais, por outro lado existem também certas vantagens de ter seu conteúdo disseminado. Os sites de companhias aéreas, por exemplo, são “vítimas” de *web scraping*. Contudo, no final de todo o processo, são elas que acabam recebendo o valor final da passagem (o valor da venda). O ponto negativo, nesse caso, é perder as informações de navegação de seus sites e o trajeto do usuário até a compra. Por outro lado, ela ganha condições de acessar novos mercados.

A atividade de *scraping* pode ser vista também como algo prejudicial. Apesar dos dados dos sites estarem disponíveis para consumo do usuário final, os proprietários dos dados alegam que eles estão disponíveis gratuitamente, mas sujeito a políticas de uso, usualmente não respeitadas por processos de *scraping*. Sendo assim, existem questões éticas acerca desta atividade, que ainda não estão bem resolvidas.

2.4 Resumo da literatura

Esta seção do trabalho dedicou-se a buscar estruturar as principais contribuições teóricas relacionadas a iniciativas de diversificação de fontes de dados em sistemas de Inteligência Analítica nas organizações. Assim, elementos como o processo decisório e a cultura analítica, bem como a estrutura de dados, disponibilidade/acessibilidade de dados e métodos de coleta foram explorados com o objetivo de suportar a presente pesquisa. A partir das referências utilizadas foi possível identificar uma série de características as quais a organização deve ter atenção, se buscar implementar uma iniciativa de diversificação de suas fontes de dados.

No que se refere ao processo decisório baseado em dados, destacaram-se elementos relacionados a cultura organizacional, bem como a importância do papel da alta liderança como instrumento de incentivo para utilização de dados no suporte à decisão. Na discussão sobre a diversificação das fontes de dados, três elementos foram destacados: Estrutura dos dados; Disponibilidade/Acessibilidade dos dados e; Métodos de coleta de dados. O primeiro item traz o debate sobre o que são dados estruturados, e destaca os principais benefícios e desafios de se trabalhar com dados não estruturados. O segundo refere-se à formação de um mercado de dados (economia de dados) cada vez mais dinâmico, com iniciativas de diferentes naturezas que interferem nas alternativas de recursos hoje disponíveis para as organizações. E por fim, o

terceiro item destaca a importância dos mecanismos de transferência de dados hoje conhecidos como *APIs/Web services*, e o papel ainda relevante das atividades de *Web Scraping*. A Tabela 2 resume os principais elementos identificados na literatura, pertinente a iniciativas de diversificação de fontes de dados nas organizações.

Tabela 2: Principais dimensões de diversidade de fontes de dados em sistemas de IA identificadas na literatura

Novas fontes de dados	Dimensão	Principais elementos	Principais autores
Processo decisório baseado em dados	Cultura	<ul style="list-style-type: none"> • Apoio da alta gestão • Intuição se sobrepondo às análises (HIPPOs) 	<ul style="list-style-type: none"> • Davenport e Harris (2007); Gudfinnsson et al. (2015); Davenport (2010); McAfee e Brynjolfsson (2012); Francisco, Kugler, & Larieira (2017)
	Dados não estruturados	<ul style="list-style-type: none"> • Infraestrutura • Ferramentas • Competências 	<ul style="list-style-type: none"> • Strauch (2011); Sullivan (2004); Chen, Chiang, & Storey (2012); Berg, Seymour, & Goel (2013); Grimes (2008); Brackett (2011)
	Dados externos à organização	<ul style="list-style-type: none"> • Compartilhamento de dados • Intermediadores/Facilitadores • Governo 	<ul style="list-style-type: none"> • Abbasi, Sarker e Chiang (2016) Hofman & Rajagopal (2014); Jetzek, Avital e Bjørn-Andersen (2013); Faria, Linden, & Laney (2017); Balazinska, Howe e Suciu (2011); Isik, Jones e Sidorova (2012); Redman (2005); Wells (2016); (Nabi, Sabir, Bilal, & Ayub, 2017); Ibragimov, Hose, Pedersen, & Zimanyi (2015); Abbasi & Adjero (2014)
	Métodos de obtenção de dados	<ul style="list-style-type: none"> • APIs e Webservices • Webscraping 	<ul style="list-style-type: none"> • Polidoro, Giannini, Conte, Mosca, & Rossetti (2015); Glez-Pena, Lourenço, López-Fernández, Reboiro-Jato, & Fdez-Riverola (2013); Chen, Chiang, & Storey (2012); Malinverno, Moyer, O'Neill, & Gilpin (2017); (Taulli, 2015); Olmedilla, Martínez-Torres, & Toral (2016); Liu e Davis (2015)
Diversidade de fonte de dados			

3 REFERENCIAL TEÓRICO

O campo de estudos de Inteligência Analítica é amplo e com muitas vertentes já pesquisadas. Assim, o conjunto de publicações que compõem sua literatura é ampla na utilização de teorias. Alguns exemplos de utilização de teorias podem ser evidenciados por autores como Li, Hsieh, e Rai, A. (2013), Wakefield (2013) e Susarla., Barua, e Whinston (2010).

O trabalho de Li, Hsieh, e Rai, A. (2013) utilizam a teoria da motivação (*Motivation Theory*) para investigar o comportamento de usuários de sistemas de informação no período de “pós-aceitação” (*post-acceptance*), no contexto do uso de ferramentas de *Business Intelligence*. Já a pesquisa de Wakefield (2013) utiliza a teoria da consistência cognitiva (*Cognitive Consistency Theory*) para explorar a influência da afetividade de usuários na divulgação de informações *online*. E por fim, Susarla., Barua, e Whinston (2010) utilizam teoria das agências (*Agency Theory*) e teoria modular (*Modularity Theory*) para estudar as interdependências entre fornecedor e cliente no contexto de soluções de *software as a service* (SaaS).

Das diferentes teorias observadas durante a revisão de literatura, a teoria da Capacidade Absortiva (*Absorptive Capacity*) mostrou-se como a de maior aderência para a condução dessa pesquisa, tendo em vista sua abordagem e finalidade. Inclusive, trabalhos como Malhotra, Gosain & Sawy (2005), Elbashir, Collier e Sutton (2011), Bharati, Zhang e Chaudhury (2014) e González e Muiña (2014) são exemplos de como a teoria encontra aderência na condução e pesquisas relacionadas a Inteligência Analítica. A concepção da teoria, bem como a descrição de seus constructos e suas utilizações serão melhor exploradas no item a seguir.

3.1 Teoria da Capacidade Absortiva

A habilidade da organização de reconhecer o valor de novas informações externas, assimilá-las e aplicá-las com finalidades comerciais, é crítica para sua capacidade de inovação. Com esta concepção em mente, Cohen e Levinthal (1990) apresentaram a teoria da capacidade

absortiva com o objetivo de identificar fatores que contribuem para o processo de inovação nas organizações. Assim, a teoria explora como as organizações se apropriam de informações externas, as combinam com informações internas, processam essas informações e transformam o conhecimento em valor para a organização.

Publicações acadêmicas ao longo da década de 90 utilizaram essa teoria como suporte em estudos com diferentes aplicações, para além da capacidade de inovação de uma organização. Temas como Sistemas Nacionais de Inovação, Aprendizado Organizacional e Gestão Estratégica compõe o conjunto de pesquisas que auxiliaram a revisão da teoria. A abordagem trabalhada por Zahra e George (2002) organiza os constructos em duas categorias de Capacidade Absortiva: Potencial (PACAP) e Realizada (RACAP). A primeira categoria diz respeito a capacidade que a empresa tem de *adquirir* informações externas e *assimila-las*, de modo a construir um conteúdo com potencial de geração de valor futuro (denominada Capacidade Absortiva Potencial - PACAP). A segunda categoria se refere a capacidade da organização de transformar essas informações e explorá-las num contexto prático, gerando valor para a organização (Capacidade Absortiva Realizada - RACAP). A Figura 5 a seguir apresenta a visão geral dos constructos propostos pelos autores.

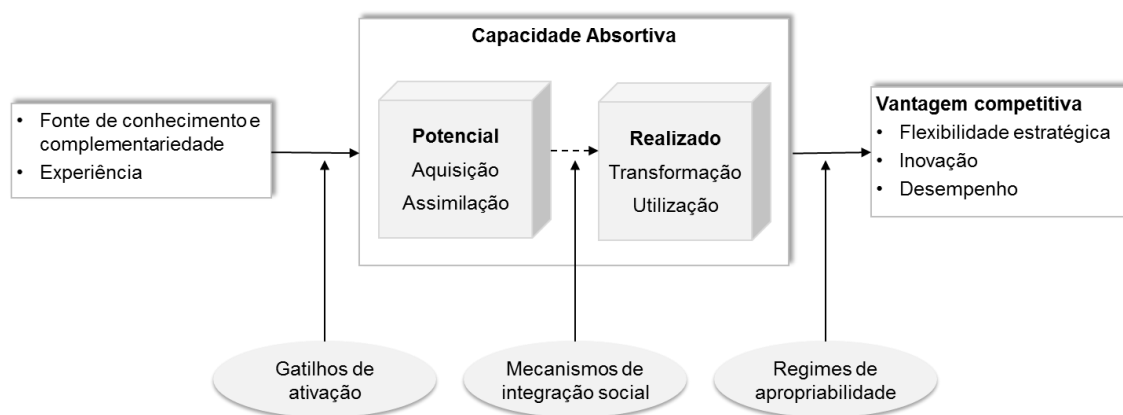


Figura 5: Capacidade Absortiva segundo Zahra e George (2002)

Fonte: Zahra e George (2002), traduzido pelo autor

No modelo teórico apresentado na Figura 5 são propostas quatro dimensões capazes de descrever a capacidade absorptiva. A *aquisição* diz respeito à capacidade de organização identificar e adquirir conhecimento gerado externamente e que são críticos para sua operação. A dimensão de aquisição pode ser melhor detalhada em função das definições oferecidas por Zahra & George (2002). Os autores sugerem que o esforço investido nas rotinas de aquisição de conhecimento possui três atributos que podem influenciar a Capacidade Absortiva:

Intensidade, Velocidade e Direção. A velocidade e intensidade dos esforços da firma para identificar e se apropriar do conhecimento pode determinar a qualidade da capacidade de aquisição da firma. Existem limites para a velocidade de aprendizado da empresa, dado que ciclos de aprendizado não podem ser encurtados facilmente, e alguns dos recursos necessários para construir a Capacidade Absortiva não são adquiridos rapidamente. O direcionamento de acúmulo dos conhecimentos também pode influenciar os caminhos percorridos pela empresa para obter novos conhecimentos externos. De forma similar, investimentos alocados anteriormente na organização, bem como o conjunto de competências instaladas, auxiliam a aumentar seu potencial de aquisição de informações.

A assimilação representa os processos e rotinas da organização que permitem analisar, processar, interpretar e compreender a informação obtida de fontes externas. A transformação descreve a capacidade da firma de desenvolver e refinar rotinas que facilitem a combinação de conhecimentos existentes e novos conhecimentos adquiridos e assimilados. Por fim a utilização, que se baseia nas rotinas que permitem à organização refinar, estender e alavancar competências existentes ou criar novas, por incorporar e transformar conhecimento nas suas operações. Para descrever os principais elementos da teoria, os autores construíram uma decomposição das dimensões de seu modelo teórico, conforme apresentado na Tabela 3, a seguir.

Tabela 3: Dimensões da Capacidade Absortiva e suas principais características, segundo Zahra e George (2002)

Capacidade Absortiva	Dimensões	Componentes	Papeis e Importância
Potencial	Aquisição	<ul style="list-style-type: none"> • Investimentos anteriores • Conhecimentos anteriores • Intensidade • Velocidade • Direção 	<ul style="list-style-type: none"> • Escopo de busca • Esquema perceptivo • Novas conexões • Velocidade de aprendizado • Qualidade do aprendizado
	Assimilação	<ul style="list-style-type: none"> • Entendimento/ Compreensão 	<ul style="list-style-type: none"> • Interpretação • Compreensão • Aprendizado
Realizada	Transformação	<ul style="list-style-type: none"> • Internalização • Conversão 	<ul style="list-style-type: none"> • Sinergia • Recodificação • Bissociação
	Utilização	<ul style="list-style-type: none"> • Uso • Implementação 	<ul style="list-style-type: none"> • Competências <i>core</i> • Colheita dos recursos

Posteriormente, Todorova e Dursin (2007) fizeram críticas à proposta de Zahra e George (2002), indicando algumas fragilidades. Entre os apontamentos, destacam-se: a falta de clareza na definição do constructo “Transformação”; a exclusão do constructo de “Reconhecimento do Valor”, proposto originalmente por Cohen e Levinthal (1990); e a agregação dos constructos em dois grupos (Potencial – PACAP e Realizada – RACAP). A revisão oferecida por Todorova e Dursin (2007) altera também as variáveis de mediação do modelo teórico, convergindo para o seguinte esquema, evidenciado na Figura 6, a seguir.

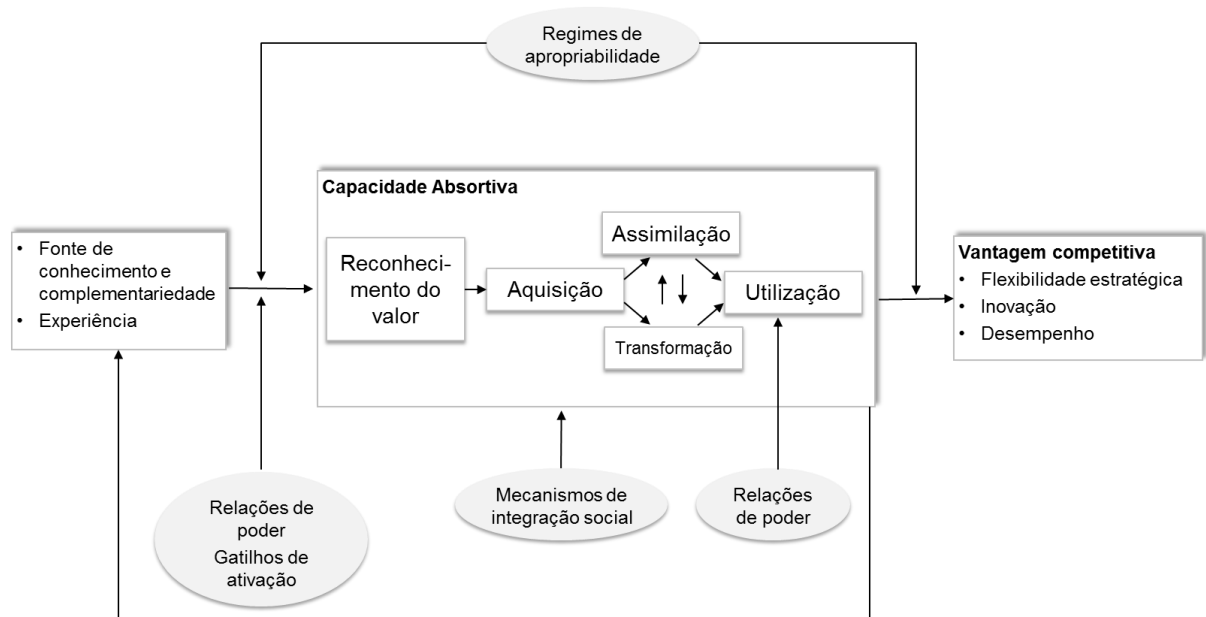


Figura 6: Capacidade Absortiva, segundo Todorova e Dursin (2007)
 Fonte: Todorova e Dursin (2007), traduzido pelo autor

Em pesquisas de Sistemas de Informação, a Capacidade Absortiva tem sido aplicada em diferentes frentes de pesquisa, tais como gestão do conhecimento, governança de TI e valor de TI para os negócios. Tradicionalmente a forma de mensurar a Capacidade Absortiva nas organizações usa *proxies* como Atividades de P&D, Estoque de Conhecimento Existente e estruturas organizacionais, rotinas e práticas de gestão de pessoas. Já em sistemas de informação, *proxies* mais comuns são Conhecimento Anterior Relacionado na firma, Inclinação para Mudanças na Gestão e Políticas de Tecnologia (Bharati, Zhang, & Chaudhury, 2014).

Para além do campo de Sistemas de Informação, as pesquisas em Inteligência Analítica também utilizam teoria da Capacidade Absortiva. O trabalho de Malhotra, Gosain e Sawy (2005), por exemplo, explora os potenciais de aprendizado organizacional ao estabelecer uma intercooperação de sistemas de informação entre a empresa e fornecedor. Elbashir, Collier e

Sutton (2011) exploram o papel da capacidade absorptiva das gerências operacionais e do alto nível de gestão na assimilação de ferramentas de BI na organização. Bharati, Zhang e Chaudhury (2014) discutem o papel da capacidade absorptiva e de pressões institucionais na assimilação das mídias sociais pelas empresas. Já González e Muiña (2014) defendem que as novas ferramentas de gestão da informação no contexto da Inteligência Analítica facilitam o desenvolvimento da capacidade absorptiva, e propõem uma revisão do conceito teórico, chamando-o de “*Smart Absorptive Capacity*” (capacidade absorptiva inteligente).

Com o objetivo de identificar como as empresas estão evoluindo seus sistemas de Inteligência Analítica para obter novas fontes de informação, este trabalho utilizará a teoria da Capacidade Absorptiva como perspectiva teórica para suportar o desenvolvimento da pesquisa. A escolha é decorrente do entendimento de que existe aderência entre a teoria e o problema de pesquisa em questão, uma vez que a teoria defende a importância de acesso ao conhecimento externo e sua interação com conhecimentos internos para gerar valor na organização. Como evidenciado, o conjunto de pesquisas que utiliza essa teoria não se dedica apenas ao processo de inovação, mas à geração de valor na empresa, num sentido mais amplo. Foram apresentados exemplos de trabalhos que utilizam essa teoria, no campo de Inteligência Analítica, reforçando a aderência da utilização da teoria na presente pesquisa.

Essa pesquisa se dedicará à exploração dos quatro constructos identificados como componentes principais da capacidade absorptiva das organizações. Estes são mencionados tanto por Zahra e George (2002) como Todorova e Dursin (2007), a saber: Aquisição, Assimilação, Transformação e Utilização, tal como evidenciado na Tabela 3. Outros elementos da teoria também poderiam ser explorados, tais como os elementos de mediação, por exemplo. Contudo, optou-se pela limitação de constructos abordados com a finalidade de estabelecer um foco de análise para a pesquisa.

4 MÉTODO

Este trabalho busca explorar quais as principais iniciativas para obtenção de novas fontes de dados em sistemas de informação de Inteligência Analítica nas organizações. Apesar de existir uma literatura extensa sobre Inteligência Analítica, a utilização de dados externos à organização, e/ou dados não estruturados nos sistemas de IA, este ainda é um tema pouco explorado neste campo de estudo. O presente trabalho pretende contribuir com este tema, buscando entender como se dão as iniciativas adotadas pelas empresas para diversificar suas fontes de dados. Deste modo, pretende-se responder a seguinte questão: quais as principais iniciativas para obtenção de novas fontes de dados em sistemas de informação de Inteligência Analítica?

Sob o ponto de vista metodológico, esse trabalho caracteriza-se como uma pesquisa de abordagem qualitativa, cujos procedimento de investigação utilizado é o Estudo de Caso, e com objetivo exploratório, segundo definições de Colauto e Beuren (2003). Ainda, a pesquisa também pode ser classificada com relação aos meios utilizados para investigação, que é a pesquisa de campo, segundo critérios definidos por Vergara (1997). A escolha do procedimento de investigação mostra-se adequada em função da natureza do fenômeno ser ainda pouco explorado na literatura (Eisenhardt, 1989). Assim, considerando a limitação de referências bibliográficas, este trabalho tem como objetivo analisar as principais características dos sistemas de IA em organizações que buscam diversificar suas fontes de dados. A escolha das empresas, bem como dos entrevistados, é dada de modo a garantir que sejam realizadas contribuições pertinentes ao campo de estudo, conforme sugerido por Stake (1994). Deste modo, os casos considerados no estudo devem ser representativos e as pessoas entrevistadas devem assumir papéis relevantes no fenômeno estudado.

Foram realizadas entrevistas semiestruturadas com o auxílio de um roteiro, conforme evidenciado no Anexo I. Todos os participantes receberam um convite por mensagem escrita sinalizando o objetivo da pesquisa acadêmica e o compromisso com o sigilo das informações. As entrevistas foram realizadas presencialmente, no ambiente de trabalho dos entrevistados, com duração média de 1 hora. Ao iniciar o encontro presencial foram entregues uma cópia do roteiro que direcionaria a entrevista (Anexo I) e uma carta de confidencialidade assinada pelo orientador dessa pesquisa (Anexo II), com o objetivo de reforçar o compromisso com o sigilo

das informações. Antes de iniciar as entrevistas foi solicitada, a todos os entrevistados, a permissão para gravar um áudio do diálogo, com o objetivo de facilitar a análise posterior do conteúdo, com transcrição e utilização de *software* para análise.

No total foram realizadas nove entrevistas com profissionais que possuem grande experiência no tema de IA, alguns assumindo os mais altos cargos de tecnologia/gestão da informação em suas organizações. Em alguns casos, em função das longas experiências profissionais, alguns entrevistados faziam referências a outras empresas que trabalharam anteriormente. Essas referências foram evitadas, mas consideradas na medida em que faziam contribuições relevantes ao tema de pesquisa em questão. A Tabela 4, a seguir, descreve algumas características relacionadas aos entrevistados que fizeram suas contribuições a esse trabalho.

Os profissionais que participaram da pesquisa possuem suas carreiras diretamente relacionadas à função de Inteligência Analítica nas organizações em que atuam. Por sua vez, as empresas onde esses profissionais trabalham possuem relevância no cenário nacional. Dois dos cinco maiores bancos do país fazem parte da amostra; duas empresas de natureza “paraestatal” de médio/grande porte, e que fazem parte do Sistema S²; e uma das maiores *startups* do país, que apesar do pouco tempo de existência, registrou em seu terceiro ano de atividade mais de dez milhões de consumidores. Todas as empresas são prestadoras de serviço e possuem unidades de Inteligência Analítica formalmente estabelecidas e com equipes estruturadas. Adicionalmente, foram considerados também os dados obtidos em entrevista com o diretor de uma consultoria especializada em projetos de Inteligência Analítica, que atende diferentes organizações, muitas de grande porte.

Os participantes da pesquisa podem ser considerados representativos, tendo em vista o perfil profissional, a capacidade de entendimento dos desafios tecnológicos associados à gestão de informações, bem como da utilização prática de dados em processos decisórios. Os entrevistados também são pessoas capazes de oferecer elementos ricos sobre o objeto estudado, uma vez que se encontram em posição adequada na organização para explorar os elementos propostos, conforme sugerido por Stake (1994).

² O Sistema S começou a ser estruturado no país em 1942 para oferecer uma rede de ensino que melhorasse a produtividade da mão-de-obra e serviços culturais e de lazer com financiamento garantido, mas sem depender da gestão pública. (Lupion, 2017).

Tabela 4: Principais características dos entrevistados

Identificador	Cargo	Tempo de experiência com IA	Setor/ Segmento da atual empresa	Número de funcionários	Tempo de existência da empresa
Entrevistado 1	Gerente de Business Intelligence - Brasil	13 anos	Banco 1	> 50.000	> 35 anos
Entrevistado 2	Gerente de TI - América Latina	8 anos	Banco 1	> 50.000	> 35 anos
Entrevistado 3	Head de Riscos Corporativos	7 anos	Banco 2	> 90.000	> 40 anos
Entrevistado 4	CDO - Brasil	15 anos	Banco 2	> 90.000	> 40 anos
Entrevistado 5	Coordenador de BI	10 anos	Sistema S 1	> 3.000	> 70 anos
Entrevistado 6	CIO - Brasil	15 anos	Sistema S 2	> 2.000	> 45 anos
Entrevistado 7	Analista Sênior	14 anos	Sistema S 2	> 2.000	> 45 anos
Entrevistado 8	Analista Sênior	5 anos	Startup	> 300	> 5 anos
Entrevistado 9	Diretor Técnico	15 anos	Consultoria (em IA)	> 20	> 5 anos

A construção do instrumento de coleta de dados levou em consideração aspectos da teoria da Capacidade Absortiva, bem como aspectos relacionados a revisão de literatura relacionada a IA. Com relação ao uso da teoria, os elementos utilizados para construção do instrumento de pesquisa estão relacionados ao apresentado na Tabela 3 deste trabalho. Sobre os aspectos de revisão de literatura considerados para a construção do instrumento de pesquisa, foram utilizados os elementos destacados conforme resumo apresentado na Tabela 2.

O resultado da síntese desses elementos pode ser verificado na Tabela 5, que apresenta a relação da teoria da Capacidade Absortiva com a revisão de literatura e o questionamento a ser feito para o entrevistado. O questionário foi submetido à análise crítica de dois especialistas da área de IA, com atuação profissional e acadêmica na área, permitindo que suas considerações enriquecessem a proposta do roteiro de entrevista. Na revisão foi dada ênfase sobre a importância de se questionar sobre a governança dos dados, considerando que nesse cenário a atuação das unidades de negócio é relevante, e a definição dos *information owners* (donos da informação) e papéis e responsabilidades torna-se um aspecto relevante. A observação foi absorvida, estando aderente a teoria da Capacidade Absortiva no constructo de *Transformação*.

Tabela 5: Roteiro construído para condução das entrevistas

Dimensões da Capacidade Absortiva	Componentes teóricos	Componentes da literatura de Inteligência Analítica	Questões propostas pelo autor
Contexto da organização (perguntas com a finalidade de identificar características gerais relacionadas ao caso estudado)			
Aquisição	<ul style="list-style-type: none"> Investimentos anteriores 	Estrutura da equipe de IA (ferramentas, infraestrutura, consultoria e pessoas)	<ul style="list-style-type: none"> Descreva de forma resumida as principais características da organização, em função do porte e participação no setor Descreva as principais características da equipe de IA e o histórico recente dessa atividade na organização
			<ul style="list-style-type: none"> Como foram trabalhados os investimentos necessários para organização passar a trabalhar com a aquisição de dados externos? (<i>Considerar aspectos como infraestrutura, consultoria, pessoas ou similar</i>)
	<ul style="list-style-type: none"> Conhecimentos anteriores 	Competências de manipulação e análise de dados	<ul style="list-style-type: none"> Para trabalhar com dados externos à organização a equipe de Inteligência Analítica necessitou de novas competências técnicas?
	<ul style="list-style-type: none"> Direção Intensidade 	Perguntas de negócio e Mindset da liderança	<ul style="list-style-type: none"> Como são definidas as necessidades de dados externos da organização? Com que frequência é feita uma revisão onde a equipe de IA se dedica a identificar e desenvolver mecanismos para captação de novas fontes de dados?
	<ul style="list-style-type: none"> Velocidade 	Evolução dos sistemas de BI&A (<i>dados externos e não estruturados</i>). Métodos de Obtenção dos dados	<ul style="list-style-type: none"> Há quanto tempo são trabalhadas as fontes de dados externos nos sistemas de Inteligência Analítica? Quais os principais métodos de coleta de dados utilizados pela organização? Existem aplicações usando APIs, web services ou Web scraping?

Dimensões da Capacidade Absortiva	Componentes teóricos	Componentes da literatura de Inteligência Analítica	Questões propostas pelo autor
Assimilação	<ul style="list-style-type: none"> Entendimento/ Compreensão 	Dados não estruturados <i>(Infraestrutura, Ferramentas e Competências)</i>	<ul style="list-style-type: none"> Que recursos são utilizados pela organização para o armazenamento e tratamento de dados não estruturados? <i>(Considerar aspectos como ferramentas de armazenamento e análise dos dados, além de competências técnicas relacionadas)</i>
Transformação	<ul style="list-style-type: none"> Internalização Conversão 	Governança de dados na organização	<ul style="list-style-type: none"> Como é trabalhada a questão da governança de dados na organização, considerando dados externos e dados não estruturados?
Utilização	<ul style="list-style-type: none"> Uso Implementação 	Dados externos à organização	<ul style="list-style-type: none"> A empresa compartilha ou recebe dados compartilhados junto a alguma organização? Quais as principais fontes de dados trabalhadas pela organização?

Os dados obtidos durante as entrevistas se resumem às anotações feitas pelo pesquisador e os áudios gravados. Os áudios foram transcritos e todos os textos foram importados para o *software* Atlas TI, que auxiliou na análise de conteúdo. O *software* é considerado uma ferramenta para a análise qualitativa de dados textuais, gráficos e vídeo. Ajuda a organizar, reagrupar e navegar em um determinado material de maneira sistemática.

O critério de categorização utilizado foi o semântico, no qual os assuntos com mesmo significado são agrupados na mesma categoria. As categorias de análise foram agrupadas conforme a coluna “dimensões” evidenciada na Tabela 3, sempre que foi possível verificar uma relação direta à temática. Caso contrário, outros códigos foram criados, e depois relacionados às categorias em questão.

O processo de tratamento dos dados aconteceu em etapas. Foi realizada uma primeira leitura, sem que houvesse a preocupação de codificação, para obtenção de familiaridade ao conteúdo das entrevistas. Na segunda leitura, em que foi feita a codificação, foram assinaladas 168 *quotations*, fragmentos de texto codificados, que foram agrupados em 29 *codes*, categorias de agrupamento dos fragmentos de texto. O detalhamento da utilização dos codes, bem como seu relacionamento com elementos da teoria da Capacidade Absortiva e principais elementos identificados na revisão de literatura podem ser observados na Tabela 6 a seguir.

Para além do conteúdo transcrito nas entrevistas realizadas, a análise também considerou conteúdos observados durante a realização das entrevistas, tal como relatórios e softwares utilizados para obtenção, tratamento e análise de dados. Em algumas entrevistas foi possível acompanhar, junto com o entrevistado, os resultados das análises propostas pela equipe de Inteligência Analítica e suas aplicações práticas na organização. Estes conteúdos foram registrados dentro das possibilidades, através de registros fotográficos ou notas realizadas pelo pesquisador. Essas medidas foram tomadas com o intuito de estabelecer uma triangulação de dados. Em algumas entrevistas foi possível aplicar a técnica observação participante, onde o observador partilha (na medida em que as circunstâncias permitam, as atividades, as ocasiões, os interesses e os afetos de um grupo de pessoas ou de uma comunidade (Argilarga, 1985).

Tabela 6: Utilização de codificação das entrevistas, e seu relacionamento com elementos teóricos

Componentes teóricos (Capacidade Absortiva)	Componentes da literatura de Inteligência Analítica	#	Code
Investimentos anteriores	Estrutura da equipe de IA. (ferramentas, infraestrutura, consultoria e pessoas)	1	Equipe
		2	Celula de Inovação
		3	Infraestrutura
		4	Investimentos
		5	Ferramenta
		6	Estrutura organizacional
		7	Evitar desperdício de recursos
		8	<i>Open source</i>
		9	Projetos
Conhecimentos Anteriores	Competências de manipulação e análise de dados	10	Competências
		11	Técnicas
Direção e Intensidade	Perguntas de negócio e Mindset da liderança	12	Direcionadores
		13	Cultura organizacional
Velocidade	Evolução dos sistemas de BI&A (dados externos e não estruturados). Métodos de Obtenção dos dados (APIs, web services e Web scraping).	14	Tempo com dados externos
		15	Métodos de coleta de dados
		16	Mecanismos não tecnológicos para obtenção de dados
Entendimento/Compreensão	Dados não estruturados	17	Dados não estruturados
		18	Fonte de dados
Internalização e Conversão	Governança de dados	19	Arquitetura
		20	Governança
		21	Hub de informações
		22	Qualidade de dados
		23	Compartilhamento
Uso e Implementação	Dados externos	24	Análise
		25	Uso dados externos
		26	Governo vendendo dados
		27	Empresas compartilhando seus dados
		28	Facebook
		29	Empresas vendendo seus dados

5 APRESENTAÇÃO DOS RESULTADOS

Os dados obtidos através das entrevistas foram analisados sob a perspectiva da teoria da capacidade absorptiva. Nos itens a seguir são apresentados os principais constructos considerados na análise e os dados associados a cada elemento da teoria, bem como seus vínculos com a literatura.

5.1 Aquisição

O constructo *aquisição* representa o esforço investido nas rotinas de aquisição de conhecimento, que devem variar em função dos seguintes elementos: Conhecimento, Investimentos, Velocidade, Direção e Intensidade. Durante a análise de conteúdo foram considerados elementos como competências para se trabalhar com Inteligência Analítica, investimentos necessários, ciclos de planejamento, visão das lideranças (inclusive de tecnologia) e experiência da equipe com trabalhos dessa natureza. Em todos os elementos foi levado em consideração o contexto de diversificação das fontes de dados da organização.

No que se refere a *competências*, existe forte convergência entre os entrevistados quanto ao perfil de profissional necessário para o desempenho de funções de Inteligência Analítica. Fundamentalmente, o profissional deve ter um perfil crítico e precisa ter uma boa visão do negócio em que atua. Conhecimentos computacionais são importantes, mas não são prioridade. Entende-se, como mais importante, que o profissional entenda o funcionamento dos sistemas de informação de uma forma geral, e não necessariamente que seja um especialista em programação de computadores. Já no contexto de obtenção de dados externos, destacaram-se habilidades de natureza interpessoal, além de uma visão sistêmica para além dos limites da própria empresa. Assim, a capacidade de identificação e articulação com outros agentes do mercado (externos à organização) aparece como algo fundamental para trabalhar com dados externos.

[Entrevistado]: É similar. Além de ser crítico e observador, tem que ter uma habilidade de articulação, né? De saber se aproximar, no caso de uma empresa. Entender onde está a informação que vai enriquecer aquilo que já existe dentro da empresa. Tem uma informação

externa, onde está? Vou dar um exemplo, no caso da [Empresa - onde o entrevistado oferece consultoria]. É importante saber quando um [cliente] morre. Onde vai buscar essa informação? É importante. Vai mandar um boleto de anuidade para alguém que já morreu? Muito desagradável para a família. E não existe o fluxo de alguém avisar que um [cliente] morreu, por exemplo. Ah, vamos buscar aqui, vamos buscar ali. Ah, mas aí você pega uma equipe que não tem esse perfil observador. Ah, isso aí não existe. Tem que pegar nos cartórios. Tem que pegar nos hospitais. Nem sempre o cara... enfim, acabamos descobrindo. Existe um instituto especialista. Responsável, que tem essa informação. Fizemos um convênio com o instituto. Saiu gratuito, o instituto não vai cobrar nada e vamos fazer esse cruzamento. Sempre é possível.

[Pesquisador]: As habilidades que mudam para se trabalhar com dados externos não é muito uma questão tecnológica, e mais uma questão de *mindset*, articulação, você abordar a outra instituição, ter uma visão mais ampla. Não só dos processos internos, mas como seu negócio interage com...

[Entrevistado]: Com o mundo, como o mundo funciona. De entender que a informação em algum lugar ela vai estar. Hoje em dia... em algum lugar está escrito. Vai estar disponível em algum lugar. Não sei onde, mas em algum lugar vai estar disponível. Se não tiver você tem que gerar a entrada do dado e dar um jeito de coletar. [...] sobretudo o perfil não é só técnico. O perfil é... da atitude, o perfil por parte do comportamento, então eu acho que gestores desse tipo de projeto, eles tem que ter uma maturidade. Maturidade de negócio. Eu entendo que não é o tipo de perfil que você consiga desenvolver muito fácil. Ele leva mais tempo para desenvolver. A pessoa tem que ter maturidade. Tem que ser uma pessoa crítica, porque ela tem que bater o olho e pensar: tem alguma coisa errada, isso aqui não está certo. E para ser crítico tem que ser observador. Tem que ter o perfil observador. Eu acho que isso é mais importante que conhecer ferramentas, sabe? **(Entrevistado 6)**

Se eu trabalho com dado, tem que ter... não precisa ser todos. Como eu te falei, o próprio [gerente] não tem uma formação tecnológica como eu tive, mas pela inteligência e habilidade dele, ele já me ouviu falar tanto da mesma coisa que ele já praticou e absorveu. Ele nunca fez um código complexo, como eu já fiz na minha vida, em três linguagens, quatro linguagens. Ele toca a vida dele e funciona. Eu tenho gente na minha equipe que é mais negócio. A grande maioria da parte *analytics* minha é codificadora mesmo. Concorde que eu não preciso ter no meu time de BI gente codificadora? Seria agradável ter gente com conhecimento lógico, raciocínio lógico, porque é a praia dessa área, dessas duas áreas. Mas não obrigatoriamente. Na hora que eu tenho um cara treinado em dados, que sabe tudo isso. Fazer a pergunta certa, pensar. Esse dado eu tenho que jogar aqui. Eu tenho que ter um *Connect:Direct*. Eu tenho que saber pedir para TI isso, entendeu? **(Entrevistado 4)**

Em geral percebe-se que as habilidades tradicionais de TI continuarão sendo importantes e que as empresas já as possuem. Quando referenciada a questão dos dados externos à organização, as habilidades tradicionais de TI já atendem satisfatoriamente às necessidades de mecanismos de coletas de dados, tal como *webservices* e APIs. Apesar de ter sofrido mudanças ao longo dos anos, esses recursos são entendidos como mais tradicionais e familiares às equipes de TI, visão que converge com o exposto por Taulli (2015). Contudo, obtenção de dados via *web scrapping* ainda se constitui como um conhecimento fora do

domínio das organizações, bem como a captura e manipulação de dados não estruturados. Apenas uma das organizações evidenciou razoável domínio sobre esses elementos.

[Entrevistado]: Mas temos também muita coisa de robô que pega em sites, assim.

[Pesquisador]: Tipo um *web scraping*, que visita página e absorve os dados?

[Entrevistado]: Tem duas situações. Tem desse tipo de coisa, usando como motor de busca o Google, puxa dado da página. Até, por exemplo, dados que ficam soltos em páginas, assim, que a gente precisa da informação. Aí você monta o robô para ele ficar pegando dados toda vez que ele atualiza lá, né?

[Pesquisador]: Você poderia ilustrar um exemplo?

[Entrevistado]: Por exemplo, o IBGE Cidades. Se você entra no IBGE Cidades você tem a produção agrícola e florestal brasileira. Você tem dados populacionais estimados das cidades. Você tem dados de registros civis. Uma série de dados lá que você consegue enxergar no IBGE Cidades. [...] Só que, se você quiser buscar esses dados na mão, você tem que entrar na cidade de Curitiba, entrar lá no ano de 2016, da produção agrícola, baixar a planilha eletrônica, aí puxar. Aí, a gente tem um robô que varre. Ele faz cidade a cidade. A gente tem de tudo. A gente tem *web service* também. A gente tem essas capturas em cima de sites. A gente tem dados de *txt* (arquivo de texto) bruto, que vem de sistemas. **(Entrevistado 5)**

API eu acho muito técnica. Muito comum, mas muito técnica. Então, eu sempre tenho equipe especialista, técnica, né. Experiência com scrapping ainda não tive. Não pedi ainda pra fazer, porque eles não sabem fazer. Eles não sabem, ainda, como garantir que aquilo está correto. É mais complexo. Então, ficamos apenas com *web service* e API, por enquanto. **(Entrevistado 6)**

A gente usa. É um complemento também. Se você perguntar a um garoto novo, que só faz isso, ele vai falar: API é o futuro de tudo. Porque ele não faz o que a gente faz. Eu acho que se você não tiver as bases você não faz API. Sem dados você não faz API. A API é muito similar ao que a gente chamava de *call* externo antigamente. Então, por exemplo, você tem API's pra jogar o CEP e saber qual o endereço completo. Essa é uma API. Então, eu acho que a ideia do API não é nova. A ideia de cientista de dados, *machine learning*, cara, isso aí mudou o nome. Você acha que, realmente, esse nome não existia? Você acha que nego inventou isso agora. Foi lá e botou duas pérolas ali e falou: agora inventei e patentei. As empresas ganham com isso, mas é a mesma coisa de antes, entendeu? Chama *Big Data*, antes era *Warehouse*. Ah, mas a tecnologia mudou. Ué, no *Warehouse* a tecnologia também mudou um monte de vezes. De DB2 pra Teradata. Vai mudando, cara, mas não muda nada. Para mim, que sou usuário, continua a mesma coisa. Então, eu acho que as API's têm um papel fundamental, como tinham antes. Uma API que eu criei você vai usar no seu programa. Vai usar no *internet banking*, vai usar na *ATM*, vai usar na agência. A mesma API você pode usar em várias coisas e ter um código rodando por trás dela. A grande vantagem da API que eu vejo, diferente do mundo anterior, eu acho que é troca de dados em diferentes plataformas, porque o resto é igualzinho, que é, por exemplo... eu chamava uma API antes, o negócio que eu podia chamar no Cobol para o Cobol. Então, se você programa em Cobol, você pode usar minha API. Só que quem usava? Se eu fiz pra Cobol você tem que usar em Cobol. Você não rodava isso em Fortran, você não rodava isso em SAS. Hoje, não. Quando você faz uma API em determinado código, você, provavelmente, vai poder usar no Java, vai poder usar no SAS, vai usar... eu acho que a grande vantagem é a portabilidade das coisas e a redução de custos. **(Entrevistado 4)**

Foi dada uma forte ênfase sobre as diferenças relacionadas à TI e Inteligência Analítica, que apesar de complementares, possuem perfis profissionais distintos. O principal aspecto alvo de críticas foi a visão de negócio, entendida como precária por parte dos profissionais de TI, inclusive considerando os níveis hierárquicos mais altos. Nesse sentido, existe uma crítica à visão de negócio das lideranças de TI e seus efeitos no negócio como um todo.

[Entrevistado]: É um problema, porque TI se preocupa com cabo de rede, com switch, com roteador. Na verdade, o cara que deveria ser o mais estratégico da empresa, é o cara mais operacional... do negócio da TI ainda. Não é nem o negócio da empresa. A maioria das TI's não sabe nem o que a empresa faz. Que é algo muito ruim. Tem todos os dados na mão, passa tudo por ele, tem acesso a tudo, mas não tem visão de negócio, né? O negócio dos caras é instalar Windows, Office, *switch*, *hub*...

[Pesquisador]: Mas você percebe isso mesmo nas posições mais gerenciais de TI?

[Entrevistado]: Mesma coisa. Poucos gerentes de TI que se envolvem na parte de negócio, estratégico. No sentido de não saber nem o que a empresa faz. A maioria. Como diz um amigo meu de uma empresa grande: TI é porteiro, a TI não pode escolher o *software* que ele vai usar. TI é igual porteiro de prédio, só tem que cuidar com quem entra (Risada). Uma forma muito radical... **(Entrevistado 9)**

Assim, eu acho que é um novo perfil. A área de TI vai continuar sendo TI. O profissional de TI é um profissional que não gosta de falar de negócio. Ele gosta de estar muito focado no que ele faz. No dia a dia dele. Se você pega equipes de desenvolvimento, o que ele quer é receber uma especificação técnica e transcrever isso num código. Essa nova era digital tem que ser pessoas comunicativas, tem que ser pessoas que entendam de negócio, tem que ser pessoas que tenham muita flexibilidade e consigam programar. Além disso, o cara tem que saber programar. Então, tem que gostar de conversar, tem que gostar de... o cara de TI não gosta de reunião, pra começar. Reunião tem que ser prática e objetiva. Fala o que eu preciso fazer e vamos fazer. Uma reunião de negócio não é assim. Você está elaborando um produto novo. É discussão, é *brainstorm*. E esse perfil novo ele entra nisso. Peraí, deixa eu ver uma coisa aqui para você. Eu vou compor esse dado, mais isso, mais isso. Eu vou agregar uma variável. O que você acha disso? Você está ali lapidando a informação. A gente tem um trabalho que é da mineração. *Data gathering*. É realmente explorar aquilo. O cara de TI não. Ele tem aquilo, ele executa aquilo. Ele não explora isso. **(Entrevistado 3)**

Na perspectiva de *investimentos* foi dada ênfase na capacitação e contratação de profissionais qualificados, mais do que qualquer outro investimento. Custos de software são mencionados, mas na visão mais ampla de sistemas de Inteligência Analítica. Assim, os custos de licenciamento são diluídos em toda a organização, e equipes que trabalham com dados externos tendem a usar a infraestrutura já disponível. No contexto da utilização de dados externos e/ou dados não estruturados, as soluções utilizadas para obtenção de dados, usualmente, são *open source* e os custos diretos estão mais associados à compra do acesso à fonte de dados, quando necessário. Apesar de existir investimento para ter acesso a determinadas fontes de dados, os entrevistados afirmam que esse não é um custo alto. Da

mesma forma, investimentos em equipamentos ficaram mais acessíveis com a disponibilidade de serviços em *cloud*, tais como aluguel de máquina virtual.

[Entrevistado]: É da ordem de uns R\$3 milhões, mais ou menos. Mas basicamente com pessoal, tá?

[Pesquisador]: A maior parte do investimento é em pessoas?

[Entrevistado]: É. Existe uma parte que eu não conseguiria nem te dizer qual o valor, que de certa forma eu fui usando infraestrutura interna, que já se tinha dentro da [Empresa]. Então, por exemplo, claro que a gente comprou um servidor poderoso para trabalhar dados, coisas dessa natureza. Mas a gente usa muito de infraestrutura instalada também da nossa TI. *Storage* e máquinas. VM's que a gente cria em cima daquilo ali. Eu não conseguiria te dar um número muito preciso. **(Entrevistado 5)**

A gente incentiva muito o pessoal a divulgar os trabalhos, colocar isso também em comunidade. A gente tem aqui um viés de trabalhar com ferramentas *open source* muito forte. A gente incentiva muito as pessoas a codificarem. Dentro do *data lake* de negócios eu tenho pagado toda a *suite* da Cloudera. Distribuição da Cloudera de *big data*. Porque a gente tem a Cloudera aqui pago, é a única peça que não é *open source*. **(Entrevistado 2)**

Eu passei três dias na sede da empresa pensando que linguagem eu vou usar. [...] na época, o [gerente] concluiu que deveríamos usar o SPSS. A gente comprou a licença. Mas aí, a pergunta é: quem usa SPSS no [empresa]? Entendeu? Aí, a grande pergunta é... eu vou usar SPSS aqui? Olhei com meu irmão. Ele está se formando em estatística. E ele trabalha com estatística. O [analista] é estatístico. Aí, fui lá. O [analista] está usando R. R com Python. Aí, você conversa com o [empresa] São Paulo usa R e Python. Aí, eu converso com meu irmão aqui, usa R e Python. Por quê? Porque é melhor. Aí, você fala assim: por que a gente comprou o SPSS? **(Entrevistado 7)**

Um dos motivos que destaca a importância dos profissionais, como investimentos críticos se dá pelo fato de que, quando se faz uma obtenção de dado externo (através de uma compra, por exemplo), é necessário ter uma capacidade interna para absorver esses dados. Adicionalmente, é necessário ter uma visão técnica para compreender onde o dado precisa ser imputado, para se obter o resultado que se espera.

Eu sou envolvido na intermediação do TI para como vai fazer funcionar. Eu sou um entregador. O que você quer fazer? Aí minha cabeça fala: esse dado tem que chegar aqui. Tem que cruzar com a base tal. Tem que colocar isso no processo produtivo. Isso vai rolar uma vez. Isso eu vou fazer o *match* dos dados e nunca mais vou fazer. Essas perguntas eu tenho que fazer para o cara. Aonde você quer chegar com isso? É *one time shot*? Eu quero fazer o enriquecimento e morreu? Ah, não deu. Quero fazer um teste. Então, essa base vai ficar durante dois meses. E depois, se der certo, a gente vai passar a comprar todo mês. Esse é o projeto. Entendeu? **(Entrevistado 4)**

De modo complementar, um dos investimentos que tende a consumir recursos são os serviços associados aos dados obtidos. É comum que fornecedores ofereçam também serviços de consultoria para as etapas de manipulação e análise dos dados. Em empresas maiores é

esperado que existam determinadas capacidades/competências instaladas, que empresas de menor porte podem não ter disponível.

O [fornecedor 1], todas as empresas, cada vez mais oferecem serviços. O que é serviço? Mão de obra. Por quê? Eles começaram a ver que desde que o mundo é mundo o [fornecedor 1] vende dado. Aí eles falaram assim: tem banco, por exemplo [empresa do setor], que não tem gente para trabalhar. Então, eu vou lá e faço para você também. Aí, o cara oferece a consultoria. Ele vem e trabalha o dado. A gente não precisa disso. Ele oferece. A [fornecedor 2] faz isso. A [fornecedor 2] é excelente como ferramenta. Eu a considero uma consultoria ruim. A [fornecedor 2] vive vendendo serviço. Acho ruim. Mas eles vivem vendendo. Por quê? Eles querem esse segmento. É mais uma maneira de abocanhar esse mercado. Deve ter demanda. De vez em quando eles vêm com um *case* aqui. Faça isso com o pé nas costas. Mas, obviamente, para alguns cegos um pouco de lanterna é luz, né? **(Entrevistado 4)**

Ainda sob o ponto de vista de investimento, há que se considerar também a característica de estrutura organizacional das diferentes empresas entrevistadas. Em uma delas existe uma equipe de aproximadamente dez pessoas dedicada a todo o processo de consumo de dados externos. Nos bancos, em ambos os casos foi mencionada a existência de equipes dedicadas a experimentação de novas tecnologias, inclusive para o consumo e manipulação de novas fontes de dados. Por outro lado, outras empresas não possuem um departamento inteiramente dedicado para o trabalho com dados externos, mas essa atividade é exercida por algumas pessoas, que podem estar posicionadas em uma ou mais unidades funcionais/administrativas.

Saiu... o Twitter lançou o Heron, que é uma ferramenta de streaming de dados. Vamos ver como esse cara funciona. Esse time fica voltado para fazer esse tipo de pesquisa. Mais do que isso, aqui dentro, não só eu, mas outras pessoas aqui também, outros gestores, que tem o papel e tem a função de transformar culturalmente a mentalidade da empresa. **(Entrevistado 2)**

A gente tem um olhar. Não na minha equipe, mas numa equipe que é parceira minha dentro da diretoria, que a gente chama de experimentação. A ideia é eles ficarem olhando isso e trazendo oportunidades de mercado. Eu não tenho tempo para isso. [...] é um colega que cuida da equipe totalmente fora da caixinha. [...] eu acho que tem coisas que as pessoas botam ali na mesa que, de cada dez, ficam dois. Sabe? E, de cada dois, um talvez seja legal, sustentável. Porque ele tem que se pagar, ser sustentável. [...] eu acho que é fundamental, inclusive está aí dentro dos *to do's* e *not to do's*. Fazer e não fazer. Dentro do fazer, cara, não perca a tecnologia. Mantenha suas *POC's* vivas, sempre olhando coisa nova e tudo. Mas você tem que entregar, cara. Se você não entregar, ninguém vai te patrocinar. **(Entrevistado 4)**

Os elementos de *Direção* e *Intensidade* foram traduzidos, nessa pesquisa, como a forma pela qual a empresa define suas necessidades de obtenção de novas fontes de dados, bem como se existem processos que formalizem ciclos de revisão dessas necessidades. Os entrevistados indicaram que não existem ciclos de planejamento previamente definidos para revisão das necessidades de novas fontes de dados. Afirmaram também que, usualmente, as necessidades

são definidas pelas unidades de negócio, que realizam demandas para a estrutura de tecnologia. Cada departamento tem sua especialidade e sua demanda. Usualmente, as unidades de negócio também utilizam seu próprio orçamento, favorecendo um modelo descentralizado. Deste modo, a forma de evolução das necessidades de dados se torna dinâmica, sem a existência de ciclos de planejamento, por exemplo.

Então, o que eu quero dizer? Cada um que cuida do seu negócio conhece isso. O cara do crédito conhece exatamente o que comprar. Ele vai no [fornecedor]. Eu fui no [fornecedor] semana passada. Tinha um cara do crédito que estava comigo, o cara da cobrança estava comigo. Eu fui com meu diretor executivo, mas tinha gente do *business* lá. Eles que, geralmente, buscam. Eu não digo nem que eles ajudam. Eles definem onde a gente deve atacar. E eles que vão pagar a conta. [...] Eu sou de uma área de dados. Normalmente, eu posso falar e dar opinião pela minha experiência? Posso. Mas a demanda geralmente vem do negócio, que conhece muito mais seu negócio do que eu. Entendeu? Eu acho que conheço bastante os negócios, mas um cara de CRM sabe muito mais da praia dele do que eu. **(Entrevistado 4)**

Em apenas um dos casos foi mencionado que todas as demandas são sistematizadas (tal como num sistema de *help desk*). Nos ciclos anuais de planejamento da empresa essas demandas são consideradas. Assim, essas informações ajudam a definir as atividades relacionadas a novas fontes de dados. Em outra empresa o coordenador das atividades relacionadas a dados externos afirmou estar, ele mesmo, sempre atento às possibilidades de uso de novas fontes de dados, para contribuir com o seu negócio. Em seu entendimento, é também papel dele sugerir e influenciar as unidades de negócios, no aumento de referências com dados externos para análises de negócio.

Assim, as pessoas fazem muitas perguntas isoladas. Ninguém faz uma pergunta estruturada. Então, o que eu montei esse ano aqui dentro. Eu fiz uma coisa que eu já tinha feito. Eu peguei o sistema de demandas e falei: vocês podem me pedir o que vocês quiserem, mas tudo é por aqui. E aí, baseado no que as pessoas me perguntam, eu comecei a entender que tipo, para que lado as pessoas estão indo. Porque você começa a olhar: tive tantas requisições indo para cá. Tantas indo para cá. Então isso aqui dá para organizar, já deixar pré-montado. E aí, uma coisa que a gente vai voltar a fazer, agora no começo do ano, é buscar quais são as perguntas-chave que as pessoas têm. **(Entrevistado 7)**

A habilidade de enxergar valor no dado é mais minha. Eu é que fico olhando e, nos diálogos com todos, né, que precisam de informação, eu vou enxergando onde é que eu posso ajudar o cara. É dali que vai me trazendo ideias para confeccionar *dashboards* ou, mesmo no diálogo com ele, vai me mostrando o que está precisando e a gente vai escovando dados para trazer entendimento para ele, né. **(Entrevistado 5)**

O constructo de velocidade se traduziu, neste trabalho, como o tempo que as equipes de inteligência analítica passaram a trabalhar também com dados externos. Em geral, as empresas

possuem equipes trabalhando com inteligência analítica há mais de dez anos, mas trabalhar com dados externos nas análises é um fenômeno razoavelmente recente. Apesar de alguns citarem o uso de dados externos há muito tempo, com a finalidade de complementação de cadastros, a utilização de outros dados (“não cadastrais”) acontece num horizonte inferior aos últimos cinco anos. Este é entendido como um fenômeno recente nas organizações, ainda utilizado de forma aquém de seu potencial e que existe muito espaço para avanço.

O dado tradicional, ERP tradicional, eu trabalho com isso aí faz uns 15 anos já. É claro que agora, ultimamente, vem crescendo. Até a parte... o pessoal começa a entender melhor essa parte de *analytics* aí, né? Essa parte de descoberta. A parte de dados externos, cara, é uma coisa de um ano e meio, dois anos para cá. Coisa nova. E a tendência é aumentar, com essa parte de análise preditiva, essa parte de *big data*, essas coisas aí. Vai começar a crescer essa informação aí. **(Entrevistado 9)**

O departamento trabalha com dados externos desde o seu nascimento, praticamente. Claro que ele era limitado a poucos dados externos, que era o que era possível acompanhar. A RAIS, por exemplo, que eram coisas mais fáceis de serem manuseadas e que tinham ferramentas. Mas você olhava informação muito agregada e era muito manual. A entrada conosco, assim, de montar um *big data*, com RAIS e mais outra coisa, um monte de fonte de dados, eu posso até numerá-las e mostrar algumas delas, é de quatro anos para cá. **(Entrevistado 5)**

5.2 Assimilação

O constructo de assimilação diz respeito à capacidade da organização de compreender os dados obtidos externamente. Tendo em vista que dados estruturados fazem parte dos sistemas de informação tradicionais, a pesquisa considerou como aspecto crítico a capacidade das organizações de assimilar dados não estruturados. Ainda, entre os dois principais aspectos estudados (origem do dado e estrutura do dado), a origem é elemento de menor complexidade para *assimilação* (no conceito proposto pela teoria da Capacidade Absortiva). Isso considerando-se que a aplicação dos dados usualmente está diretamente relacionada ao negócio, o que favorece alguma familiaridade prévia dos analistas com os dados. A familiaridade aumenta se esses dados são do tipo estruturado. Nesse sentido, buscou-se identificar que elementos relacionados aos dados não estruturados estão sendo trabalhados nas organizações, considerando aspectos de infraestrutura, tecnologias e ferramentas de armazenamento, manipulação e análise, assim como competências técnicas associadas

Os dados coletados sugerem que existe uma lacuna a ser preenchida nas organizações. Apenas dois casos foram apresentados de forma concreta. Um deles exemplifica a utilização de dados não estruturados em análise de *logs*, como parte de um processo de verificação de operações financeiras. O outro diz respeito a uma captura de dados de notícias, realizando a compilação dos dados para uma unidade funcional/administrativa consumir. Os outros exemplos mencionados pelos entrevistados foram apresentados como em fase de projetos, mas sem especificações sobre a escolha de ferramenta ou tecnologias de armazenamento e manipulação dos dados. A falta de padronização das tecnologias foi um dos aspectos mencionados que dificultam o trabalho com esse tipo de dado, convergente com a visão de Strauch (2011).

Internamente nós temos pouca experiência com o tratamento de dados não estruturados. Enfim, a gente faz algumas análises em cima dos *logs* da nossa aplicação. E esse dado não é estruturado. Dá certo trabalho para extrair a informação. Nós temos umas regras internas. Até comentei a questão da Fazenda, de validar o CPF. Então, se o CPF é inválido ou a pessoa, na regra do negócio, tem menos de 16 anos, a gente barra a negociação antes de tomar qualquer ação sobre ela. Então, ela fica somente no *log* hoje. A gente não chega a criar um identificador que vai identificar algum pagamento. Do lado da [cliente], foi uma requisição que ela tentou fazer para gente, mas que, do nosso lado, não foi algo processado. E aí, a gente tem que comparar um por um, o que eles mandam e o que a gente recebe. A gente tem que fazer essa análise dos *logs* também. Para poder comparar os números corretamente. **(Entrevistado 8)**

Tudo que a gente está falando aqui é [sobre] dados estruturados. Mas, com dados não estruturados, eu ainda não sei se a gente tem estrutura para trabalhar. Eu estou estudando um pouco, à parte daqui, a questão de como trabalhar. Inclusive, montei ali dois servidores Python. Uma máquina que eu tenho ali para começar a brincar, entender oficialmente, como trabalhar com dado não estruturado. Mas é muito pouco isso ainda. É muito incipiente. Por quê? São tantas fórmulas, mas não é uma fórmula bem montada. E além de não ser uma fórmula bem montada, um programa da vida que a gente instala nesse momento, eu vejo que, assim, ele faz pra responder uma pergunta específica. E as pessoas acham assim: que joga e ele vai trazendo informação. Na verdade você tem que trazer a pergunta, e modelar essa pergunta. **(Entrevistado 7)**

[Entrevistado]: Pra estruturado, ainda a gente usa muito o SQL Server, da Microsoft. É a ferramenta padrão, aqui dentro da [empresa]. Para alguns semiestruturados, assim, que vem em XML, a gente usa ainda o próprio SQL Server, porque ele tem suporte para isso. E a gente achou mais conveniente continuar com ele. Tem bons motores de busca. Melhoraram muito nas últimas versões, ele tem bom suporte para isso. A gente também tem uma, que eu até esqueci o nome agora, mas a gente tem muita informação agora que... a gente tem uma base de dados em específico, ela vem sempre numa cadeia de caracteres, ela não segue muito o padrão, então a gente também está usando o SQL estruturado para esse fim, porque aí você tem motores de busca muito rápidos em cima de campos com grandes cadeias de caracteres e milhões de linhas. Aí quando cai no não estruturado, que são os *documents*, aí a gente usa

MongoDB. O primeiro a gente estava usando aqui, localmente, configurando a estrutura toda com essa finalidade. Mas, aos poucos, a gente já migrou para a nuvem.

[Pesquisador]: Há quanto tempo vocês estão usando essa experiência com o MongoDB?

[Entrevistado]: Um ano, mais ou menos. É, que é um mundo muito complexo também, sabe. As tecnologias, tem muitas, distintas tecnologias, é difícil decidir. É difícil de administrar. Então, assim, você tem que usar nuvem. Você gasta muito tempo. Quando a gente estruturou o MongoDB para funcionar, ainda estava desenvolvendo, estava testando, tal, o volume ainda era pequeno. Porque quando você entra no mundo NoSQL, a sua escalada em termos de dados é muito maior. O estruturado eu ainda consigo, quando vem uma nova base que eu vou trabalhar, alguma coisa que a gente vai fazer, a gente já sabe quanto vai crescer. O NoSQL, como você fica varrendo distintas coisas, já não tem exatamente esse parâmetro. Você não sabe, não consegue. Então, eu achava melhor a gente migrar para a nuvem. Só que eu não queria sair já gastando na nuvem. Enfim, nem era tão caro o investimento, só que eu não queria sair gastando, sem saber quando que eu ia entregar o produto, sabe. Eu achava melhor a gente... pode me chamar de muquirana, mas foi isso. Nem é tão caro. Para você ter uma ideia no último mês deu US\$ 400, US\$ 500 por mês. Se considerar, não é caro. Mas, enfim, aí a gente fez localmente. Gastamos um tempo administrando, fizemos localmente. E aí eu consultei os caras da Microsoft antes. Nós estamos fazendo assim, tal, tal, tal. Na hora que eu for subir para nuvem tem alguma dificuldade? Não, tranquilo. Cara, na hora que subiu para nuvem não rodava nada. E era só transposição de dados para lá. Não era nem levar a aplicação, era só fazer o dado ir para lá. Porque tinham funções de agregação que faltavam no Azure, coisas dessa natureza. Então, é complicado. **(Entrevistado 5)**

Um dos entrevistados mencionou a dificuldade de se definir o que são dados não estruturados. Disse discordar da convenção, segundo a qual, são dados que fogem do padrão relacional, conforme apresentado na literatura deste trabalho. Apesar de não oferecer uma definição concisa, ele resgata a visão de que são dados não modelados, convergente com Grimes (2008). Nas palavras do entrevistado, dados não estruturados possuem argumentos não padronizados, o que dificulta muito sua manipulação. Por fim, fez associação direta dos dados não estruturados com textos, sons e imagens.

[Entrevistado]: É, existem várias pessoas que dizem que dado não estruturado é um dado que não está organizado. Dado que não é relacional. Existem várias coisas. Eu acho que não é isso, tá? Eu fiz dois filmes para mostrar pra pessoas do banco, leigas, que você vai entender o que eu estou falando. [...] O que eu quero te mostrar aqui? [...] o que demanda uso de ferramentas de dados não estruturados, para mim são dados que não tem uma definição comum, por exemplo, um atributo. Quando você fala em dados não estruturados começa a ter e poder usar dados como, por exemplo, uma imagem, um texto. Um texto do Facebook, um comentário do Facebook. Como é que você trabalha com isso? Não é cruzando. Você tem que ter uma linguagem interpretativa. Provavelmente essa linguagem vai pegar algumas coisas, algumas pesquisas de opinião, que pega, assim... o cara está feliz com esse produto. Análise comportamental. É outro tipo de pegada. [...]

[Pesquisador]: E aqui, vocês já chegaram a ter alguma experiência de usar CouchDB, MongoDB ou similares?

[Entrevistado]: Nem sei o que é isso.

[Pesquisador]: Algum processador de dados não estruturados...

[Entrevistado]: Não, nem sei o que é isso. O que eu conheço, o que a gente busca dentro do *big data*, a gente usa Hadoop, na tecnologia Horton Works. Partição *data lake*, área de usuário, tudo a gente está trabalhando. Ainda não está funcionando para analista, está funcionando só para produção, um anexo do banco que está lá. **(Entrevistado 4)**

Outro entrevistado mostrou ceticismo com relação às aplicações de dados não estruturados existentes hoje nas organizações. Mencionou que, em essência, os dados trabalhados se resumem a telemetrias, *logs* e similares. Ao mesmo tempo, não acredita na “promessa” inicial de que dados não estruturados brutos (como texto, imagem etc.) consigam ser analisados sem a devida estruturação/modelagem. Adicionalmente, o esforço de trabalhar com dados não estruturados conflita com o esforço de trabalhar com os dados já estruturados, que nem sempre se encontra em estágio maduro nas organizações.

Eu não acredito nisso. Tanto é que na primeira versão de *big data* que os caras falavam, só conversa. Pegar dado não estruturado do Word, Excel, mas não é realidade. Hoje não faz parte da realidade. Tanto é que *big data*, hoje 95% da execução é só para leitura de *log*... por quê? Porque tem muito volume de dado. Quando você liga o celular ele gera um monte de *log*. Que é isso aí que os caras estão... que nem hoje, tem muita coisa de agronegócio, que é a parte de telemetria de equipamento, de colheitadeira, de extrator, não sei o quê, que gera todo *log* de posição de máquina, tudo. Toda telemetria na máquina, lê cada segundo. Aí usa *big data* para isso. E não vai existir. Não tem como. O cara não consegue cuidar nem do dado estruturado dele. **(Entrevistado 9)**

5.3 Transformação

O constructo de *transformação* representa a capacidade da organização de internalizar dados externos aos seus sistemas de informação. Essa competência foi traduzida neste trabalho como a governança dos dados, buscando explorar, especialmente, aspectos relacionados a novas fontes de dados.

Foi possível observar que as empresas possuem razoável facilidade na absorção de informações de natureza cadastral, como mencionado anteriormente. Esta facilidade se deve ao fato de existirem campos-chave bem definidos e de definição universalizada (como CPF e/ou CNPJ). Esses dados são comuns às fontes internas e externas, o que facilita muito a integração de dados dessa natureza. Contudo, para outros dados externos não foram mencionados exemplos explícitos de integração de dados, exceto em uma das empresas. Nesse caso

específico foi mencionado que, neste ano, foi feito o primeiro exercício de integração dos dados internos e externos diretamente na plataforma de Inteligência Analítica da organização.

A gente vem desenhando, cada vez mais... os primeiros experimentos aconteceram neste ano, mas tudo indica que nós vamos aumentar e cruzar cada vez mais esses dados. Já existem algumas leituras nesse sentido e elas vão aumentar. Isso depende muito... nós aqui da [empresa] ainda dependemos de cada área, assim. Tecnicamente falando, nós estamos preparados para isso, tanto em termos de arquitetura, como em alinhamento entre as equipes. Da TI nossa. Mas as áreas de negócio, ainda, vamos dizer, estão engatinhando nesse sentido, mas elas estão, aos poucos, utilizando cada vez mais dessas informações. **(Entrevistado 5)**

Os entrevistados mencionaram a importância de se estabelecer uma governança de dados na organização. Porém, alguns apresentam suas ideias ainda em fase de projetos, sem mencionar aspectos práticos de governança. Três entrevistados fizeram observações objetivas, de como buscam trabalhar elementos de governança de dados. Os principais pontos explorados são o desenvolvimento da cultura de uso dos dados corporativos e das possibilidades de uso dos dados externos.

Mas isso passa muito, esse cruzamento de dados internos e externos, passa muito por uma mudança cultural da organização. Eu percebo, sabe. Primeiro, eles nunca tinham visto tanta informação simultânea, externa. Então, assim, só de ele conseguir enxergar uma situação já é para ele um choque, que ele tem que digerir e ver o que que ele vai fazer com aquilo ali. Aí você dá logo uma pancada... tem que ir aos poucos. Tem uma mudança cultural que ocorre. **(Entrevistado 5)**

Um dos elementos que aumenta a complexidade de iniciativas de governança e, por sua vez, aumenta a dificuldade de integração dos dados corporativos com dados externos, é a autonomia das unidades de negócio na contratação dos dados que irão utilizar. As múltiplas contratações em diferentes unidades de negócio acabam por criar silos de informação, sendo um estímulo para o desenvolvimento de uma governança de dados, como ferramenta para diminuir esse efeito.

[Entrevistado]: Então, eu imagino que, para a metade de 2018, mais ou menos, o [empresa] pode falar assim: estou maduro e estou utilizando *big data* de verdade, tá? BI já utilizamos muito aqui dentro. Tem 300 tipos de ferramentas de BI aqui dentro. Cada área de negócio trata seus negócios como BI. O cara tem servidor de baixo da mesa e coloca o que ele quer dentro do BI. Então, nós somos responsáveis pelo quê? Fazer gestão dessas ferramentas gerais. Riscos tem um, CRM tem outro, cobrança tem outro. É distribuído. Por isso quando você perguntou: que tecnologia que você usa? Cara, aqui usamos tudo. O gestor de negócios fala assim: quero usar o Tableau. Você tem dinheiro para pagar o Tableau? O Tableau é caríssimo. Ah, eu vou pagar, porque eu tenho dinheiro aqui para pagar licença do Tableau. O cara sabe onde é a dor dele. Ele sabe o que tem que fazer para atingir as metas dele e cumprir os objetivos. Então, cada unidade de negócio tem o seu. E, assim, uma das nossas missões aqui, hoje, é virar todos esses BI's, que estão todos espalhados aí, e centralizar todas essas

informações dentro do *data lake*. Por exemplo, tem Exadata aqui, Exadata da Oracle, extremamente caro. O que eu consigo processar no Exadata eu consigo fazer o mesmo processamento dentro do *data lake*. Qual o problema? Quem é que vai converter aqueles PL/SQL que estão aí... que o cara nem sabe mais o que faz, mas está lá rodando, para uma linguagem de *spark*, por exemplo. Então, esses são alguns pontos que estão sendo discutidos e vão sendo levados em frente, para conseguir também deixar o [empresa] atendendo todos e ter a uniformidade, trabalhando todos com os mesmos dados, conceito de comunidade, de replicar a informação dentro de micro serviços. São todas coisas que a gente pensa com os dados do *lake*. Isso é um dos pontos. Que também a governança de todos esses dados dentro do Lake e que outras pessoas vão usar. É outro ponto de preocupação que a gente tem. A gente sente falta de algumas ferramentas, para essa governança de dados.

[Pesquisador]: E o [empresa] chega a discutir quando vai contratar dados externos...

[Entrevistado]: Algumas conversas sim, mas fica limitado, não na utilização do dado, mas nas integrações técnicas que vão ser feitas. Para a conversa com diversos fornecedores, somos chamados para a gente entender. Se atende, não atende. Saber se o cara não está enganando com questões tecnológicas. A gente fica ali como advogado, um escudo de tecnologia. Os dados são deles. Todos os dados que a gente administra aqui dentro do *lake* são deles. Então, eles têm que ter autonomia mesmo.

[Pesquisador]: E o que vem de dado externo vocês acabam nem administrando?

[Entrevistado]: Eles montam lá ou planilhas de Excel. Se viram. O trabalho é jogar isso para dentro do *data lake*.

[Pesquisador]: Mas isso é um interesse seu?

[Entrevistado]: Meu e do [empresa].

[Pesquisador]: Esses silos de informação... a ideia é diminuir?

[Entrevistado]: É eliminar. **(Entrevistado 2)**

No contexto da utilização de dados externos, quando esses dados convergem para os sistemas de Inteligência Analítica, usualmente não se utilizam do fluxo conceitual de informações normalmente difundido na literatura (Sistemas de operação > Core Data Warehouse/Data Mart > Sistemas de Análise, tal como descrito por Baars e Kemper (2008) na Figura 1). Por vezes, as unidades de negócio acabam por criar seus próprios bancos de dados e sistemas para suportar suas atividades. Neste cenário, os dados externos acabam fazendo um desvio da etapa de convergência para um *Data Warehouse*, passando por um banco de dados paralelo ou se dirigindo diretamente à camada analítica.

[Entrevistado]: Você pode ver que é engraçado, mas dados externos [...] a TI não se apropriou disso até hoje. Então, tudo que precisa de dados externos, inclusive eu tenho gerado muito material para imprensa. Isso começou a funcionar até. As pessoas começaram a olhar a gente como um *hub* de informação. Mas é tudo a gente que gera.

[Pesquisador]: Isso nem chega a passar pelo DW?

[Entrevistado]: Não. A gente arrumou direto. Está tudo aqui. Ou estão no banco ou no *qvd* [camada de *data mart* na ferramenta *Qlikview*]. Agora montamos um banco organizado, paralelo. Então está funcionando. Tem usuário, tem senha...

[Pesquisador]: E como é a governança disso? Não tem muita interferência da TI com isso?

[Entrevistado]: Zero. **(Entrevistado 7)**

Outro aspecto ressaltado como recorrente nas empresas é a utilização de dados externos em momentos diferentes do uso dos dados internos. Elementos como credenciais de acesso, diferentes etapas do processo decisório e distinções de funções departamentais contribuem para que os dados externos sejam analisados num contexto e os dados internos em outro.

Então, por exemplo, o cara vai lá e tem toda informação de IBGE, CAGED, RAIS, Secretaria de Fazenda, Secretaria de Agricultura, essas coisas. Ele tem esses dados, só que são isolados do mundo dele. Então, outra ideia que a gente tem é tentar misturar esses dados com os dados do negócio dele. Por exemplo, se a gente pegar uma base, que é a agrícola... Insumos agrícolas. O cara vai lá. Ele tem mercado em Maringá, por exemplo. Ele está vendendo produto para algodão e milho. Só que, se ele for consultar na secretaria de agricultura lá, ele vai ver que em Maringá o foco é soja. O que acontece? Ele está desperdiçando esforço no negócio. Então, a ideia seria a gente juntar essa informação da agricultura, falando que ali é um polo de soja, e ver o que ele está focando, forçando para ele vender insumo para aquele...[...] O que acontece? Esses dados externos vêm de um monte de lugares. Você fazer um DW aí seria uma coisa meio óbvia, pelo sentido de você juntar informação, facilitar essa montoeira de dados, informações que vêm de um monte de lugares. Mas ele ainda é isolado do negócio. Ele passa num DW, mas o DW... os dados não conversam. Geralmente, o cara é que sabe. O cara vai, ele tem noção do negócio dele. Ele tem a parte de análise do negócio. Ele tem, mais ou menos, a imagem do cenário que quer ver. E daí, ele vai no dado externo, para pesquisar aquele cenário. Mas, as informações poderiam ser mais fáceis. Poderia fazer o cenário junto. Poderia ter um cenário atual, cenário de realização, de ver o que aconteceu, vendo já uma perspectiva de um cenário externo, como é que ele estaria. [...] normalmente, tem um departamento. Isso que eu falei com você. Até no conceito de processo é separado. Você tem um departamento que é de... como é que vou chamar, um genérico aí? Mercado. Dados de mercado, essas coisas. O cara trabalha com esses dados. E ele nem tem acesso à informação do negócio deles. Aí já tem um gap, que é reflexo do dado embaixo, lá. O cara faz uns estudos de mercado, que na verdade é análise de mercado, com esses dados aí. Toma diretiva com dado totalmente externo e passa para parte comercial, para o cara aplicar...
(Entrevistado 9)

5.4 Utilização

O constructo de *utilização* refere-se à capacidade da organização de explorar os conhecimentos externos como um recurso capaz de gerar valor. Essa visão foi utilizada no trabalho, considerando o cenário de compartilhamento de dados entre empresas, por meio de produtos e serviços, além da consequência de diversificação das fontes de dados nas organizações.

Todos os entrevistados mencionaram o uso de informações cadastrais, como forma de higienização e enriquecimento de seus dados internos, bem como atribuições de escalas de risco

para pessoas físicas e jurídicas. Deste modo, esse é um mercado relativamente tradicional, com concorrentes estabelecidos e diferentes alternativas de contratação disponíveis para as organizações. Para além da utilização de dados cadastrais, as empresas classificadas como “paraestatais” indicaram maior utilização de dados externos, com iniciativas já finalizadas em ferramentas de inteligência analítica e disseminadas na organização.

O banco tem muita informação e trabalha muita informação. Quando eu pego assim: que dado externo a gente trabalha? Vai ser um [fornecedor] da vida, um *bureau* de mercado, alguma coisa assim. Isso é informação externa ao banco, mas é informação que está no banco há 20 anos, há 30 anos. **(Entrevistado 1)**

A RAIS é uma informação boa que a gente está usando também. O CAGED a gente está usando para o número de funcionários. [...] as empresas que a gente atendeu, elas aumentaram o número de funcionários? Diminuíram? Então, esses são tipos de coisas que eu queria começar a fazer mais, mas eu preciso de tempo. Nesse ponto, eu espero fazer mais isso agora. **(Entrevistado 7)**

Então, como nós somos uma área de pesquisa, a gente, desde o princípio, só teve olhares para dados externos. O departamento existe há 13, 14 anos. Ele trabalha com dados externos desde o seu nascimento, praticamente. [...] Aí tem Ministério do Trabalho e Emprego, aí tem Educação, aí tem fontes de micro dados, de Ministérios, dados abertos, públicos, do Governo, mas tem também muita coisa de robô, que pega em sites, assim. [...] IBGE Cidades. Dados que são publicados anualmente. Dados de finanças públicas municipais, de todos os municípios brasileiros. A gente tem de tudo. [...] esses dados, por exemplo, do IBGE, que eu citei. Eles ainda são dados estruturados. Agora, a gente tem também o não estruturado, quando ele fica varrendo a internet atrás de coisas, aí já vem um dado NoSQL, mesmo. **(Entrevistado 5)**

Contudo, todas as empresas mencionaram exemplos de iniciativas em andamento e/ou possibilidades futuras de serem trabalhadas algumas fontes de dados externos de forma sistemática.

Para uma decisão, uma tomada de parecer, de um empréstimo pra uma empresa que, de repente, tem vinculação do nome dela num tipo de trabalho escravo, isso é um problema, porque aí eu tenho vinculação de imagem. Então, quando a gente fala de risco, o arcabouço de risco é muito grande. São N riscos. [...] a gente já faz, mas faz de uma maneira manual. Uma pessoa analisando, capturando informação. E fazendo um parecer na mão. Quem sabe, no futuro, a gente possa capturar isso de uma maneira mais estruturada, massiva, e cruzar com outras coisas. Que é um pouco do que o mercado está evoluindo, né? Todo mundo começa a falar que algumas ações que são feitas, realizadas por pessoas, num futuro vão deixar de existir. **(Entrevistado 3)**

Por exemplo, uma troca que estava sendo iniciada. A gente fez um convênio com uma *startup*. A *startup*... foi um edital, na verdade. Essa *startup* que obedeceu aos critérios e acabou ganhando, pra fazer o convênio. Uma *startup* de ERP para empresas, gratuito. Qual é o projeto? Que a partir da interação e de um documento, um termo de compromisso de confidencialidade, assinado, dessa *startup* com a [empresa], poderíamos usar os dados dos

usuários dos sistemas. Contas a pagar, contas a receber. Era o ERP, tinha PDV. Era para restaurantes especificamente. Depois que ela ampliou. As informações que as empresas lançam nesse ERP, o propósito que a gente já estava desenvolvendo isso era construir, a partir do *business intelligence*, você construir inteligência das informações. E saber na prática, quais são os principais *gaps* e dificuldades das empresas, no que tange à gestão. **(Entrevistado 6)**

Os tipos de fontes de dados variam entre bases de dados públicas (IBGE, MTE, MDIC etc.), outras empresas do mercado (intituladas de “empresa parceira”), bem como investimento em serviços de dados transacionados de acordo com condições contratuais. No caso das empresas “paraestatais”, foi mencionado um “acordo de cooperação técnica”, celebrado entre entidades estatais, normalmente, com o objetivo de fortalecimento entre as instituições. Sob esses acordos as instituições trocam dados detalhados de suas operações e estatísticas oficiais. Por exemplo, o Ministério do Trabalho, através desses acordos, pode disponibilizar dados detalhados das empresas que contrataram e demitiram funcionários durante um ano.

Aí aconteceu um problema, porque eles cancelaram o contrato, eles pediram, justamente pra não ter mais problema, um acordo de cooperação técnica. Então, a gente parou de ter, por algum tempo, esse acesso. Eu continuo com acesso a FTP lá dentro do MTE, onde eu busco todas as bases. Não identificadas, inclusive, ajustada, não ajustada. [...] então, já tem isso. Só que eles cortaram a identificação por esse período. Mas, passou inclusive pela diretoria nacional, agora, faz uma semana. A aprovação do acordo de cooperação. Está pra voltar já. Aí, voltando, a gente volta a ter informação identificada, com CPF, CNPJ, inclusive com salário, com tudo. **(Entrevistado 7)**

É, a gente compra essa informação da Receita. Tem que validar CPF, ver a data de nascimento. Enfim, fazer essa validação. Por exemplo, pra pessoa não estar utilizando um CPF inválido. Isso é um serviço que a gente precisa comprar, que tem interface com a Receita. Fora isso, a gente também tem um serviço de antifraude. É um serviço que analisa, basicamente, comportamento e o histórico de um usuário na internet e quando ele bate no [empresa] a gente utiliza as informações que essa aplicação gera, para determinar um *score* de risco. Para as compras com cartão de crédito. Aí não é de CPF. É de um ID de usuário que navegou nas páginas. Enfim, é um conjunto de dados. De onde ele está comprando, valor de compra, que cartão que ele está usando. [...] Acho que as informações que ela possui para analisar são limitadas. Mas, enfim, tudo funciona dentro de uma *machine learning*. Então, a gente vai retroalimentando todas as informações. É bem parceira nossa, inclusive. Então, a gente passa boa parte das nossas transações para eles analisarem e a gente indica algumas que foram fraudes no passado, que acabou dando suporte no que a ferramenta é hoje. É uma *startup*. Se chama [nome da empresa mencionada]. É de São Paulo. Enfim, tem várias ferramentas antifraude no mercado, mas a similaridade de ser *startup*... temos maior afinidade. **(Entrevistado 8)**

A empresa que possui um departamento totalmente dedicado a exploração de dados externos já apresenta análises sofisticadas e já integradas à plataforma de Inteligência Analítica

corporativa. Algumas das análises consideram, inclusive, os dados internos da organização, mesclado com dados externos, de forma simultânea na análise.

[Enquanto mostrava um dos *dashboards* da empresa] o gráfico explica como você analisa os cursos. Então, cada quadrante desse aqui representa o posicionamento desse curso dentro da sua regional. Se ele estiver dentro dessa linha para cá, para a direita, quer dizer que ele está bem posicionado, em termos técnicos. E o que é bem posicionado em termos técnicos? Como é que está a tendência do mercado de educação técnica daquele curso, ou seja, como é que ele está tendendo em termos de crescimento ou decrescimento na região. Como é o nosso posicionamento frente à concorrência. Qual é a participação privada daquele curso, naquela regional. Como é que nós, [empresa], isso aqui já é um dado interno, como é que nós, [empresa], crescemos ou decrescemos naquele mercado regional. O quanto que nossas matrículas subiram ou caíram. E, principalmente, como é que é a demanda de trabalho por aquele técnico. Aqui tem três fontes de dados externas e uma interna. A nossa interna é a própria execução de alunos matriculados. As externas são três basicamente: dados de educação do INEP, do censo educacional do país, de nível básico. O censo da educação básica pega a educação infantil, fundamental e de ensino médio, aí entra técnico junto. A RAIS Trabalhadores, para ver o estoque de mão de obra, porque aí nós criamos uma taxonomia em cima, para ligar. Quando você está falando de um técnico de mecânica [...] cada tipo de curso tem ocupações [profissionais] associadas. E mais o CAGED, que daí a gente vê a flutuação de demanda do mercado de trabalho por aquele profissional. Então se a demanda do mercado de trabalho está desaquecida, ela está em queda, significa que, naturalmente... isso já é um trabalho de cruzamento de dados, mas naturalmente a tendência de educação técnica do mercado, para aquele profissional, cai também. Mesmo porque vai ter gente sobrando no mercado. Aí tem outro detalhe. Se a gente continuar induzindo a formação daquele tipo de profissional, a gente vai formar um cara que não vai encontrar um mercado de trabalho. Aí, consequentemente, nós vamos desperdiçar recurso. **(Entrevistado 5)**

Nos bancos entrevistados percebeu-se uma postura um pouco mais conservadora, com relação ao uso de dados externos. Em síntese, existe uma maior preocupação com o ajuste dos dados internos, desafio que já possui bastante complexidade, por si. Assim, apesar de existirem equipes dedicadas à identificação de inovações tecnológicas, inclusive com esse olhar de dados externos, elas parecem estar desconectadas das necessidades de negócio. Os exercícios de acompanhar as inovações tecnológicas parecem estar mais direcionados ao entendimento da tecnologia em si, do que em pensar em como aquilo pode agregar para o negócio. Por outro lado, percebe-se certo receio de bancar o risco da iniciativa não oferecer resultados imediatos à organização. É mencionado que pode ser melhor, em alguns casos, esperar que um fornecedor externo desenvolva soluções dessa natureza e a comercialize já num estágio mais maduro.

Aqui no banco eu tenho várias operações para tocar, para manter, então chega alguém com uma ideia superinovadora. Vamos pegar dados do Twitter e fazer uma análise de sentimentos sobre o que os clientes estão falando, como, por que, o que está agradando ou desagradando. Ok. Eu faço isso ou eu melhoro meu caixa de autoatendimento? Então, concorre isso com "n" projetos que, às vezes, você não tem fôlego de andar e vem outra empresa de fora e fala assim:

já fiz, está aqui, e você precisa disso por conta disso, disso, disso e me vende a solução. Às vezes é mais fácil. Aqui é difícil de ter pessoas que pensem tão fora da caixa assim. **(Entrevistado 1)**

Talvez seja... o banco, nesse sentido, todas as organizações financeiras, elas são muito mais reservadas. Do que uma empresa de telecomunicações, que você tem que, realmente, penetrar no mercado, porque a migração do cliente para outra operadora é muito grande. Banco já não tem isso, né? Ele é um pouco mais conservador. Pode ser, assim, que até tenham áreas, como o marketing, que trabalha um pouco mais de mídia social. E o [empresa], agora, está incorporando a parte do banco digital. Talvez ele passe a olhar isso. Mas, de novo, o banco tem várias frentes. O banco é composto de “n” organizações bancárias, securitárias e não bancárias, comerciais. Em algum momento isso pode vir a convergir, mas quando a gente fala de banco, ele se restringe um pouco mais... **(Entrevistado3)**

6 CONCLUSÕES

Por meio das entrevistas e levantamento de dados foi possível verificar que as empresas, de fato, estão buscando alternativas de fontes de dados para evoluir suas plataformas de Inteligência Analítica. Seja através da utilização de fontes de dados externos à organização e/ou dados não estruturados, os entrevistados estabeleceram diálogo com a temática proposta e fizeram considerações direcionadas ao seu contexto de atuação profissional. O estudo de caso múltiplo se mostrou um método aderente à pergunta de pesquisa e rico para o desenvolvimento do trabalho, na medida em que conseguiu capturar diferentes realidades aplicadas ao mesmo elemento teórico estudado.

A escolha da perspectiva teórica da Capacidade Absortiva se mostrou aderente à proposta de pesquisa e auxiliou na análise, na medida em que ofereceu elementos que permitiram decompor o processo de obtenção e utilização de dados externos. Os constructos utilizados ofereceram diretrizes para a realização de análises específicas, tais como investimentos, competências, governança, entre outros.

Foi possível observar que o uso de dados externos, num sentido amplo, não é algo propriamente novo nas organizações. Considerando todos os sistemas de informação, notam-se aplicações específicas já realizadas há muito tempo pelas empresas. Entre elas, é possível destacar a higienização e enriquecimento de informações cadastrais, assim como a atribuições de perfil de risco a pessoas físicas e jurídicas (chamadas, no jargão do sistema financeiro, como “*scoring*” – derivação da palavra inglesa *score*).

O uso de dados externos, em sistemas paralelos aos sistemas de informação formais, também é uma prática relativamente antiga e muito difundida. Essa utilização usualmente acontece através de contratações feitas diretamente nas unidades administrativas (unidades de negócio) da empresa. As unidades, usualmente, possuem bastante autonomia para decisão sobre contratações de serviços de dados, sem uma preocupação específica com a convergência das informações junto aos sistemas formais de informação. Essa prática contribui para a consolidação de silos de informação e, por vezes, tende a estar associada a atividades manuais de manipulação de dados.

A convergência dos dados demandados nas unidades administrativas para os sistemas formais de informação é uma escolha, que pode ou não ser feita, e depende bastante da visão

das lideranças de tecnologia na organização. Em uma das empresas, observa-se que o departamento responsável pela consolidação dos dados externos não é a equipe de TI. Em outra empresa, foi indicado que a TI tem um olhar atento à questão de governança de dados, e que estuda possibilidades para estabelecer processos capazes de organizar esse tipo de dado. As demais organizações não sinalizaram iniciativas de governança, considerando essas fontes de dados, e atribuem acentuada autonomia às unidades de negócio, em termos de orçamento, soluções administrativas e tecnológicas.

A utilização dos dados externos, nos sistemas de Inteligência Analítica, tem ampliado seu escopo num período relativamente recente (últimos cinco anos). Para além do uso tradicional de dados cadastrais, as empresas começam a incorporar, em suas análises, algumas bases oriundas do governo e empresas atuantes no mercado. A utilização de bases de dados de Ministérios da União Federal, empresas privadas que comercializam e/ou compartilham dados, foi mencionada em diferentes entrevistas. Em função do movimento de utilização de dados externos em sistemas de Inteligência Analítica ser relativamente novo, alguns entrevistados sinalizaram certa expectativa de crescimento do uso dessas informações no futuro.

Os mecanismos de coleta de dados, tais como APIs e *web services*, são muito difundidos e aproveitam competências tradicionais, alocadas na equipe de tecnologia da organização. Já a prática de *web scrapping* não aparece como uma competência detida pelas empresas, sendo objeto de confusão ou dificuldade de compreensão nas entrevistas. Apenas em um dos casos o entrevistado soube apresentar uma aplicação real, bem como desenvolveu um discurso eloquente relacionado ao tema.

Dados não estruturados foram outro elemento indicado como fora do domínio de conhecimento dos entrevistados. Essa constatação converge com dois aspectos mencionados na literatura: existe uma necessidade de maior difusão de conhecimentos relacionados a dados não estruturados nas matrizes curriculares de profissionais da área de tecnologia (Fowler, Godin, & Geddy, 2016) e; necessidade de padronização/convergência das soluções de manipulação de dados não estruturados, tal como no caso de SQL/bancos relacionais (Strauch, 2011).

Os investimentos para diversificar as fontes de dados são baixos, uma vez que essa prática consegue se utilizar dos recursos já existentes. Os custos diretos de obtenção dos dados, bem como máquinas e *softwares* associados, não foram considerados, pelos entrevistados, como algo significativo. Foi indicado também que, quando existe a necessidade de novos recursos de software, é muito comum existir soluções *open source* disponíveis, principalmente

utilizando linguagens como R e Python. Os principais elementos de investimento indicados pelos entrevistados são pessoas com competência adequada (considerando aspectos técnicos/tecnológicos e visão de negócio), bem como eventuais consultorias necessárias, dependendo da estrutura organizacional.

A dinâmica de necessidade de dados externos está muito vinculada às demandas das unidades de negócio, normalmente tratadas caso a caso. Nesse sentido, observa-se que as empresas não exercem uma prática de planejamento da obtenção desses dados, tendo em vista as necessidades da empresa como um todo. Apenas um dos entrevistados informou sistematizar as demandas recebidas pelas unidades com a finalidade de utilizar essas informações como racional para um planejamento relacionado à necessidade de dados externos.

A governança de dados torna-se um elemento mais complexo nas organizações, quando considerados os dados de fontes externas e/ou não estruturados. A principal característica atenuante dessa questão é a forma como as empresas se organizam. Em geral, as unidades administrativas possuem papel de central importância no processo de aquisição dos dados externos. Da mesma forma, o posicionamento da equipe de TI é de certa passividade nesse contexto. Assim, existem soluções de gestão de dados paralelas aos sistemas formais e difundidas nas organizações. Essas características parecem aumentar a complexidade da estruturação de um projeto de governança de dados em escala corporativa.

Esse conjunto de observações, levando em consideração investimentos, competências, direcionamento, integração e governança de dados, bem como sua utilização, apresentam uma visão geral de como o uso de dados externos e/ou não estruturados na organização deve influenciar diferentes fatores. Tendo em vista todos esses elementos, a partir da teoria da capacidade absorptiva é possível elaborar uma síntese, a partir dos principais elementos associados à diversificação de fontes de dados, conforme a Figura 7, apresentada a seguir.

Capacidade Absorativa			
Potencial		Realizada	
Aquisição	Assimilação	Transformação	Utilização
<ul style="list-style-type: none"> Conhecimentos de tecnologia tradicional são necessários (<i>webservice</i>, manipulação de dados estruturados etc.) Perfil profissional deve ser questionador e com habilidade interpessoal capaz de estabelecer relações para além dos limites da organização As organizações utilizam dados cadastrais de fontes externas há algum tempo, mas a utilização de outras fontes de dados como forma de complementar análises do negócio é um fenômeno recente (menos de 5 anos) Investimentos na compra direta de dados, softwares e hardwares são indicados como inexpressivos. O principal investimento está associado a mão de obra relacionada a essa atividade. Mecanismos de coleta de dados tal como APIs e <i>web services</i> são bastante difundidos, enquanto <i>web scraping</i> apresenta pouca experiência prática 	<ul style="list-style-type: none"> As organizações apresentam dificuldades de manipulação e interpretação de dados não estruturados Aplicações existentes são limitadas em algumas organizações, em geral a análises de <i>logs</i>, telemetrias e similares Maior parte das empresas entrevistadas não possuem projetos relacionados a dados não estruturados 	<ul style="list-style-type: none"> A integração de dados externos e internos possui maior maturidade no contexto dos dados cadastrais Para outros dados externos a maioria das organizações não possuem processos bem definidos e governança adequada para utilização dos dados internos e externos de forma consolidada Usualmente as organizações fazem uso de dados externos de forma dispersa em diferentes unidades funcionais. Na perspectiva de TI não existe um entendimento claro se essa questão é algo que deve ser tratado. 	<ul style="list-style-type: none"> Empresas caracterizadas como “paraestatais” apresentaram maior familiaridade com o uso de dados externos Os <i>bureaus</i> de crédito e empresas que comercializam dados cadastrais são as fontes de dados externos mais mencionadas nas empresas Foram observadas trocas de informações sem o envolvimento de montantes financeiros nas instituições “paraestatais” e <i>startups</i> Uma das organizações pôde evidenciar análises relativamente complexas, utilizando dados externos e internos simultaneamente. As análises ficam disponíveis em ferramenta de Inteligência analítica e são utilizadas para suportar decisões em nível tático na organização.

Figura 7: Síntese dos principais aspectos relacionados a diversificação das fontes de dados
Fonte: elaborado pelo autor

Considerando os diferentes perfis de organização analisadas neste trabalho, bem como os dados obtidos a partir dos entrevistados, é possível associar às organizações diferentes estágios de desenvolvimento/maturidade, com relação à diversificação de fontes de dados utilizados em seus sistemas de informação de Inteligência Analítica. Esses diferentes estágios podem ser observados em função da utilização de dados externos, dados não estruturados e métodos de coleta de dados utilizados.

Em síntese, foi possível notar que, no contexto de dados externos, as organizações possuem maior experiência com a utilização de dados cadastrais e de *scoring* de risco financeiro, que pode ser entendido como uma prática amplamente difundida e relativamente antiga. Outras fontes de dados externos possuem utilização relativamente restrita a algumas organizações e são práticas recentes. De forma análoga, dados não estruturados possuem uma

categoria, cuja utilização é relativamente bem difundida, que são os dados de telemetria, *logs* e similares. Outros dados não estruturados são pouco utilizados, mesmo que em estágio de projeto ou iniciativas em estágios iniciais. Por fim, os mecanismos de coleta de dados mais difundidos estão relacionados às APIs e *webservices*, enquanto a prática de *web scraping* ainda é bastante limitada nas organizações. Uma síntese dessas considerações pode ser observada na Figura 8 a seguir.

	Dados externos		Dados não estruturados		Mecanismos de coleta de dados
	Dados Cadastrais	Outros	Dados de telemetrias, logs e similares	Outros	
Estágio de desenvolvimento da organização					
Baixo	<ul style="list-style-type: none"> • Não usa • Realiza cargas em lote • Usa de forma integrada nos sistemas da organização 	<ul style="list-style-type: none"> • Não usa • Usa em diferentes unidades funcionais, de forma dispersa • Possui processos bem definidos para gestão de dados obtidos de fontes externas 	<ul style="list-style-type: none"> • Não usa • Utiliza para gerenciamento de algum processo operacional da organização 	<ul style="list-style-type: none"> • Não usa • Possui projetos para testar ferramentas e infraestrutura • Possui uma aplicação que suporta decisões 	<ul style="list-style-type: none"> • Não realiza • Coleta dados em lote (via FTP e similares) • Possui aplicação para APIs e <i>web services</i> • Possui coleta de dados via <i>web crawlers</i>
Alto					

Figura 8: Estágios de desenvolvimento da organização na diversificação de fontes de dados
Fonte: Elaborado pelo autor

As conclusões da pesquisa possibilitaram a identificação dos principais elementos relacionados ao desafio de diversificação das fontes de dados nas organizações. Assim, tanto os resultados obtidos, como a revisão literatura, ajudam a construir uma visão geral sobre o tema, oferecendo a pesquisa como uma contribuição para a literatura relacionada. Contudo, a pesquisa possui limitações como o volume de entrevistas e empresas envolvidas no estudo, o que dificulta realizar generalizações para um universo mais amplo de empresas. Ainda, outras abordagens metodológicas e lentes teóricas podem contribuir para a melhor compreensão deste fenômeno em trabalhos futuros. Ou ainda, a utilização da mesma teoria, mas considerando todos os constructos teóricos mencionados por Zahra e George (2002) ou Todorova e Dursin (2007), deve trazer elementos para complementar a visão explorada por esse trabalho. Outras oportunidades de pesquisa podem estar associadas a elementos específicos do contexto de

diversificação de dados. Tais elementos podem ser entendidos como a difusão da utilização de APIs nas organizações, aspectos éticos relacionados à troca de informações entre as organizações, monetização dos dados oriundos do governo e o papel de tecnologias emergentes (a exemplo das mídias sociais, *blockchain*, *IoT* etc.), por exemplo. Adicionalmente, a partir de algumas falas dos entrevistados foi possível notar que os fornecedores de dados (e soluções associadas a esses dados) podem ser atores de importante interlocução nas próximas pesquisas relacionadas a esse tema.

REFERÊNCIAS

- Abbasi, A., & Adjero, D. (2014). Social media analytics for smart health. *Trends & Controversies*, 14.
- Abbasi, A., Sarker, S., & Chiang, R. H. (2016). Big data research in information systems: toward an inclusive research agenda. *Journal of the Association for Information Systems*, 17(2). Acesso 20 jun. 2017, em <http://aisel.aisnet.org/jais/vol17/iss2/3/>.
- Albertin, A. L. (2001). *Comércio eletrônico: modelo, aspectos e contribuições de suas aplicações* (3. ed.). São Paulo: Atlas.
- Anderson, C. (2008). The end of the theory: the data deluge makes the scientific method obsolete? *Edge*. Acesso 15 jun. 2015, em http://www.edge.org/3rd_culture/anderson08/anderson08_index.html.
- Anguera Argilaga, M. T. (1985). Metodología de la observación en las Ciencias Humanas (3ª ed rev.). Madrid: Cátedra.
- Baars, H., & Kemper, H. G. (2008). Management support with structured and unstructured data: an integrated business intelligence framework. *Information Systems Management*, 132-148.
- Balazinska, M., Howe, B., & Suciu, D. (2011). Data markets in the cloud: an opportunity for the database community. *Proceedings of the VLDB Endowment*, 4(12), 1482-1485. Acesso 20 jul. 2017, em <http://www.vldb.org/pvldb/vol4/p1482-balazinska.pdf>.
- Berg, K. L., Seymour, T., & Goel, R. (2013). History of databases. *International Journal of Management & Information Systems*, 17(1).
- Bharati, P., Zhang, C., & Chaudhury, A. (2014). Social media assimilation in firms: investigating the roles of absorptive capacity and institutional pressures. *Information Systems Frontiers*, 257-272.
- Bose, R. (2009). Advanced Analytics: opportunities and challenges. *Industrial Management & Data Systems*, 155-172.

- Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, 15(5), 662-679. Acesso 25 jul. 2017, em <https://doi.org/10.1080/1369118X.2012.678878>.
- Brackett, M. (2011). What are unstructured data? *Dataversity*. Acesso 20 maio 2017, em <http://www.dataversity.net/what-are-unstructured-data/>.
- Burn-Murdoch, J. (2012, December 19). Study: less than 1% of the world's data is analysed, over 80% is unprotected. *The Guardian*. Acesso 20 maio 2017, em <https://www.theguardian.com/news/datablog/2012/dec/19/big-data-study-digital-universe-global-volume>.
- CanalTech (2017). *O que é API?* Acesso 20 jul. 2017, em <https://canaltech.com.br/o-que-e/software/o-que-e-api/>.
- Chalk, S. J. (2015). Leveraging Web 2.0 technologies to add value to the IUPAC solubility data series: development of a REST style website and application programming interface *API*, 87, 1127–1137.
- Chen, H., Chiang, R. H., & Storey, V. (2012). Business intelligence and analytics: from big data to big impact. *MIS Quarterly*, 36(4), 1165-1188.
- Chou, S. W. (2005). Knowledge creation: absorptive capacity, organizational mechanisms, and knowledge storage/retrieval capabilities. *Journal of Information Science*, 31, 453–465.
- Cohen, W. M., & Levinthal, D. A. (1990, March). Absorptive capacity: a new perspective on learning and innovation. *Administrative Science Quarterly*, 35, 128-152.
- Colauto, R. D., & Beuren, I. M. (2003). Coleta, análise e interpretação dos dados. Como elaborar trabalhos monográficos em contabilidade: teoria e prática, 3, 117-144.
- Comunidade Transparência Hacker. (2011). *Manual dos dados abertos: governo*. [S. l.]: Laboratório Brasileiro de Cultura Digital e Núcleo de Informação e Coordenação do Ponto BR.
- Davenport, T. H. (2006). Competing on analytics. *Harvard Business Review*, 84-98.
- Davenport, T. H. (2010). *Inteligência analítica nos negócios: como usar a análise de informações para obter resultados superiores*. [S. l.]: Elsevier.

- Davenport, T. H., & Harris, J. I. (2007). *Competing on analytics: the new science of winning*. [S. l.]: Havard Business School.
- Debortoli, S., Müller, O., & Vom Brocke, J. (2014). Comparing business intelligence and big data skills. *Business & Information Systems Engineering*, 6(5), 289-300.
- Demchenko, Y., Grosso, P., Laat, C., & Membrey, P. (2013, May). Addressing big data issues in scientific data infrastructure. *Procceding of the International Conference on Collaboration Technologies and Systems, San Diego, California*, pp. 48-55.
- Eisenhardt, K. M. (1989). Building theories from case study research. *The Academy of Management Review* , 532-550.
- Elbashir, M. Z., Collier, P. A., & Sutton, S. G. (2011). The role of organizational absorptive capacity in strategic use of business intelligence to support integrated management control systems. *The Accounting Review*, 86(1), 155–184.
- Faria, M., Linden, A., & Laney, D. (2017, June 16). Understand the data brokerage market before choosing a provider. *Gartner*. Acesso 10 jul. 2017, em <https://www.gartner.com/doc/3746017/understand-data-brokerage-market-choosing>.
- Ferrández, A., Maté, A., Peral, J., Trujillo, J., Gregorio, E. D., & Aufaure, M. (2016). A framework for enriching data warehouse analysis with question answering systems. *Journal of Intelligence Information Systems*, pp. 61–82.
- Fonseca, C., Meirelles, F., & Diniz, E. (2010). *Tecnologia bancária no Brasil: uma história de conquista, uma visão de futuro*. [S. l.]: Ciab Febraban .
- Fowler, B., Godin, J., & Geddy, M. (2016). Teaching case: introduction to NoSQL in a traditional database course. *Journal of Information Systems Education*, 27.
- Francisco, E. d., Kugler, J. L., & Larieira, C. L. (2017). Líderes da transformação digital. *GV Executivo*, pp. 23-26.
- Freitas, H., Oliveira, M., Saccol, A. Z., & Moscarola, J. (2000). O método de pesquisa Survey. *Revista de Administração*, 105-112.
- Gilad, T., & Gilad, B. (1986). Business intelligence: the quiet revolution. *Sloan Business Review*.

- Glez-Pena, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2013). Web scraping technologies in an API world. *Briefings in Bioinformatics*, 15(5), 788-797.
- Goes, P. B. (2014, September). Editor's comments: big data and its research. *MIS Quarterly*, 38(3).
- González, P. M., & Muiña, F. E. (2014). Absorptive capacity and smart companies. *Intangible Capital*, 922-947.
- Grimes, S. (2008, August 1). Unstructured data and the 80 percent rule. *Breakthrough Analysis*. Acesso 18 abr. 2017, em <https://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/>.
- Gudfinnsson, K., Strand, M., & Berndtsson, M. (2015). Analyzing business intelligence maturity. *Journal of Decision Systems*.
- Hofman, W., & Rajagopal, M. (2014). A technical framework for data sharing. *Journal of Theoretical and Applied Electronic Commerce Research*, 9(3), 45-58.
- Holsapple, C., Lee-Post, A., & Pakath, R. (2014). A unified foundation for Business Analytics. *Decision Support Systems*, pp. 130-141.
- Ibragimov, D., Hose, K., Pedersen, T. B., & Zimanyi, E. (2015). Towards exploratory OLAP over linked open data: a case study. In M. Castellanos, U. Dayal, T. B. Pedersen, & N. Tatbul, *Enabling real-time business intelligence*. [S. l.]: Springer.
- ICT. (2016). *ICT facts and figures*. [S.l.: S. n.]
- IDC. (2012). *IDC digital universe study: big data, bigger digital shadows and biggest growth in the far east*. [S. l.]: EMC Corporation.
- IDG. (2016). *IDG enterprise data and analytics survey*. [S. l.: S. n.]
- Instituto Brasileiro de Defesa do Consumidor - Idec. (2017). *Aplicativos do Grupo Pão de Açúcar desrespeitam Marco Civil da Internet*. Acesso 17 nov. 2017, em <https://idec.org.br/noticia/aplicativos-do-grupo-pao-de-acucar-desrespeitam-marco-civil-da-internet>.

- Isik, O., Jones, M. C., & Sidorova, A. (2012, January). Business intelligence (BI) success and the role of BI capabilities. *Intelligent Systems In Accounting, Finance and Management*, 18, 161–176.
- Jetzek, T., Avital, M., & Bjørn-Andersen, N. (2013). Generating value from open government data. *Proceeding of the International Conference on Information Systems, Milan, Italy*, 34. Acesso 18 jul. 2017, em <https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1181&context=icis2013>.
- Kelly, J. (2014). The data economy manifesto. *Wikibon*. Acesso 08 maio 2017, em http://wikibon.org/wiki/v/The_Data_Economy_Manifesto.
- Kimble, C., & Milolidakis, G. (2015). Big data and business intelligence: debunking the myths. *Global Business and Organizational Excellence*, 23–34.
- Kiron, D., & Shockley, R. (2011). Creating business value with analytics. *MIT Sloan Management Review*, p. 57.
- Kugler, J. L. (2013). *Competência analítica*. São Paulo: Saraiva.
- LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Business Review*, 42(2).
- Li, X., Hsieh, J. P. A., & Rai, A. (2013). Motivational differences across post-acceptance information system usage behaviors: An investigation in the business intelligence systems context. *Information systems research*, 24(3), 659-682.
- Liu, P. H., & Davis, M. E. (2015). Web scraping: limits on free sample. *American Bar Association*.
- Luhn, H. P. (1958). A business intelligence system. *IBM Journal*.
- Lupion, B. (2017, fevereiro 18). O que é o Sistema S, quanto custa e a quem beneficia. *Nexo*. Acesso 22 jul. 2017, em <https://www.nexojornal.com.br/expresso/2017/02/18/O-que-%C3%A9-o-Sistema-S-quanto-custa-e-a-quem-beneficia>.
- Malhotra, A., Gosain, S., & Sawy, O. A. (2005). Absorptive capacity configurations in supply chains: gearing for partner-enabled market knowledge creation. *MIS Quarterly*, 145-187.

- Malinverno, P. (2016, February 19). The API economy: turning your business into a platform (or your platform into a business). *Gartner*. Acesso 10 jul. 2017, em <https://www.gartner.com/doc/3217617/api-economy-turning-business-platform>.
- Malinverno, P., Moyer, K. R., O'Neill, M., & Gilpin, M. (2017, January 25). Top 10 things CIOs need to know about APIs and the API economy. *Gartner*. Acesso 10 jul. 2017, em <https://www.gartner.com/doc/3579034/top--things-cios-need>.
- March, S. T., & Hevner, A. R. (2005). Integrated decision support systems: a data warehousing perspective. *Decision Support Systems*, 1031–1043.
- McAfee, A., & Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard Business Review*.
- Meirelles, F. S. (2017). *Pesquisa anual do uso de TI*. São Paulo: FGV.
- Miguel, P., & Ho, L. (2010). Levantamento tipo survey. In *Metodologia de pesquisa em engenharia de produção e gestão de operações* (pp. 71-128). Rio de Janeiro: Campus.
- Murphy, M. (2016, maio 21). The rise of APIs. *Tech Crunch*. Acesso em 20 set. 2017, em <https://techcrunch.com/2016/05/21/the-rise-of-apis/>.
- Nabi, Z., Sabir, N., Bilal, M. A., & Ayub, N. (2017). A comparative study of data federation tools for integration. *International Journal on Information Technologies & Security*, 9(1).
- Najork, M. (2009). *Web crawler achitecture*. [S.l]: Microsoft Research.
- Natis, Y. V. (2016, July 12). Hype cycle for platform as a service, 2016. *Gartner*. Acesso 10 jul. 2017, em <https://www.gartner.com/doc/3371726/hype-cycle-platform-service->.
- Negash, S. (2004). Business intelligence. *The communications of the Association for Information Systems*, 13, 54.
- OKF. (2012). Open Knowledge Foundation. *Open Data handbook*. Acesso 09 nov. 2016, em: <http://opendatahandbook.org/guide/en/>.

- Olmedilla, M., Martínez-Torres, M. R., & Toral, S. (2016). Harvesting big data in social science: a methodological approach for collecting online user-generated content. *Computer Standards & Interfaces*, 46, 79–87.
- Open Data Institute. (n.d.). *The data spectrum helps you understand the language of data*. Acesso 17 jul. 2017, em <https://theodi.org/data-spectrum>.
- Petrini, M., & Pozzebon, M. (2009). Managing sustainability with the support of Business intelligence: integrating socio-environmental indicators and organisational context. *The Journal of Strategic Information Systems*, 178-191.
- Polidoro, F., Giannini, R., Conte, R. L., Mosca, S., & Rossetti, F. (2015). Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation. *Statistical Journal of the IAOS*, 165–176.
- Power, D. J. (2008). Understanding data-driven decision support systems. *Information Systems Management*, 149–154.
- QuandL. (2017). Acesso 01 jul. 2017, em <https://www.quandl.com/about>.
- Randall, L., & Beyer, M. (2015, October 05). Strategic roadmap for modernizing your data warehouse initiatives. *Gartner*. Acesso 10 jul. 2017, em <https://www.gartner.com/doc/3142720/-strategic-roadmap-modernizing-data>.
- Redman, T. C. (2008). *Data driven: profiting from your most important business asset*. [S. l.]: Harvard Business Review.
- Rodrigues, T. (2017). Na prática: dados alternativos em processos anti-fraude. *Bigdatacorp*. Acesso 02 jul. 2017, em <https://www.bigdatacorp.info/single-post/2017/02/20/Na-Pratica-Dados-Alternativos-em-Processos-Anti-Fraude>.
- Sadiq, S., & Indulska, M. (2017). Open data: quality over quantity. *International Journal of Information Management*, 37(3), 150–154..
- Schoenherr, T., & Speier-Pero, C. (2015). Data science, predictive analytics, and big data in supply chain. *Journal of Business Logistics*, 1(36), 120–132.
- ScrapeSentry. (2015). *The scraping threat report*. Acesso em 15 set. 2017, em <https://www.scrapesentry.com/the-scraping-threat-report-2015>

- Stake, R. E. (1994). Case studies. In N. K. Denzin, & Y. S. Lincoln, *Handbook of qualitative research*. Thousand Oaks, CA: Sage.
- Strauch, C. (2011). *NoSQL databases*. [S.l.: S. n.].
- Struckman, C., & Spencer, M. (2017, February 09). 2017 CIO agenda: a U.S. perspective. *Gartner*. Acesso 15 jul. 2017, em <https://www.gartner.com/doc/3597441/-cio-agenda-perspective>.
- Sullivan, D. (2004). Text mining in business intelligence. In M. Raisinghani, *Business intelligence in the digital economy: opportunities, limitations, and risks* (pp. 98-110). [S. l.]: Idea Group.
- Susarla, A., Barua, A., & Whinston, A. B. (2010). Multitask agency, modular architecture, and task disaggregation in SaaS. *Journal of Management Information Systems*, 26(4), 87-118.
- Taulli, T. (2015, november 30). Why the API economy is exploding. *Forbes*. Acesso 21 jul. 2017, em <https://www.forbes.com/sites/tomtaulli/2015/11/30/why-the-api-economy-is-exploding/#32f21c4261e9>.
- Teodoro, A. N., Przybilowicz, É., & Cunha, M. A. (2014). Governança de tecnologia da informação: uma investigação sobre a representação do conceito. *Revista de Administração*, 44(2), 307-321.
- Todorova, G., & Dursin, B. (2007). Absorptive capacity: valuing a reconceptualization. *Academy of Management Review*, 774–786.
- Vedder, R. G., Vanecek, M. T., Guynes, C. S., & Cappel, J. J. (1999). CEO and CIO perspectives on competitive intelligence. *Communications of the ACM*, 108–116.
- Vergara, S. (1997). Metodologia de pesquisa social. São Paulo: Atlas, 44-50.
- Viri, N. (2017, julho 28). Pão de Açúcar descobre um tesouro nos algoritmos. *Brazil Journal*. Acesso 29 jul. 2017, em <http://braziljournal.com/pao-de-acucar-descobre-um-tesouro-nos-algoritmos>.
- Wakefield, R. (2013). The influence of user affect in online information disclosure. *The Journal of Strategic Information Systems*, 22(2), 157-174.

- Wells, P. (2016). Widening the debate on data sharing. *Open Data Institute – ODI*. Acesso em 09 jul. 2017, em <https://theodi.org/blog/widening-the-debate-on-data-sharing>.
- Willcocks, L., Venters, W., & Whitley, E. (2014). *Moving to the cloud corporation: how to face the challenges and harness the potential of cloud computing*. [S. l]: Palgrave Macmillan.
- Zahra, S. A., & George, G. (2002). Absorptive capacity: a review, reconceptualization, and extension. *Academy of management review*, 27(2), 185-203.

ANEXO I

Objetivo

A pesquisa busca entender que iniciativas as organizações estão adotando para evoluir os sistemas de suporte a decisão (Business Intelligence/Analytics etc.), neste trabalho denominado de Inteligência Analítica (IA). Mais especificamente, explora como as empresas estão trabalhando fontes de dados externos à organização e dados não estruturados.

Descrição Geral

- Descreva de forma resumida as principais características da organização, em função do porte e participação no setor
- Descreva as principais características da equipe de IA e o histórico recente dessa atividade na organização

Estudo de caso

- **Há quanto tempo** são trabalhadas as fontes de dados externos nos sistemas de Inteligência Analítica? Durante a evolução das atividades, que desafios puderam ser percebidos?
- Quais **as principais fontes de dados** trabalhadas pela organização?
- *Como é trabalhada a questão da governança de dados na organização, considerando dados externos e dados não estruturados? (Existe alguma consideração sobre governança dos dados? Quem são os information Owners? Isso é importante no momento da aquisição de dados?)*
- **Como foram trabalhados os investimentos necessários** para organização passar a trabalhar com a aquisição de dados externos? *(Considerar aspectos como infraestrutura, consultoria, pessoas ou similar. Quais foram os critérios para definição destes investimentos?)*
- Para trabalhar com dados externos à organização, a equipe de Inteligência Analítica necessitou de **novas competências técnicas**?
- Como **são definidas as necessidades de dados externos** da organização?
- **Com que frequência é feita uma revisão** onde a equipe de IA se dedica a **identificar** e desenvolver mecanismos para **captação de novas fontes** de dados?
- **Que recursos** são utilizados pela organização para o **armazenamento e tratamento de dados não estruturados**? *(Considerar aspectos como ferramentas de armazenamento e análise dos dados, além de competências técnicas relacionadas)*
- A **empresa compartilha ou recebe dados compartilhados** junto à alguma organização?
- Quais os principais métodos de coleta de dados utilizados pela organização? **Existem aplicações usando APIs, web services ou web scraping?**

ANEXO II

Carta de confidencialidade entregue para os entrevistados, indicando o nome da empresa, mês e ano de referência, e assinada pelo orientador da pesquisa.



Prezados senhores da [Empresa],

A entrevista sobre o uso de novas fontes de dados em sistemas de *Business Intelligence & Analytics* faz parte de uma atividade de pesquisa em andamento na FGV-EAESP, com o objetivo de trazer para reflexão questões relacionadas a dados externos à organização e seu uso nas empresas.

Os dados desta pesquisa serão compilados e analisados por alunos de pós-graduação (em particular João Gabriel Saraceni) da Fundação Getúlio Vargas (FGV/EAESP). **As informações individuais prestadas para esse estudo são totalmente confidenciais** e serão usadas apenas de forma agrupada, como insumo para um estudo de caso, de modo a não permitir a identificação da empresa e seus representantes.

Contamos com a sua valiosa colaboração e desde já, agradecemos a participação.

São Paulo, [mês] de [ano].

[Assinatura]

Prof. Dr. Fernando S. Meirelles
Coordenador da Pesquisa do GVcia